Summer 8-15-2021

Does the Combination of Spacing and Testing Promote Transfer Beyond Either Strategy Alone?

Zeynep Oyku Uner
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological & Brain Sciences

Dissertation Examination Committee:
Henry L. Roediger, III, Chair
David Balota
Andrew Butler
Mark McDaniel
James Wertsch

Does the Combination of Spacing and Testing Promote Transfer Beyond Either Strategy Alone?

by

Oyku Uner

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

Graduate school is undoubtedly challenging, but I am fortunate to have known many wonderful people that made the journey to this point a rewarding one. First, I want to thank Roddy Roediger for his mentorship throughout graduate school. Roddy is an incredible advisor. He has always offered a refreshing perspective to my many questions, he has taught me how to write and how to tell a good story, he has been very supportive and respectful of my career goals, he is eager to learn no matter how much he already knows, and he makes people around him better. I learned a lot from Roddy about being a good academic, and I will always see him as my mentor.

I would also like to thank my dissertation committee members—Mark McDaniel, Andy Butler, Dave Balota, and Jim Wertsch. Their feedback during the dissertation process was invaluable, and I learned greatly from each one of them outside of my dissertation. Clearly, the dissertation is not the work of just one person. I need to thank my labmate and officemate Wenbo Lin for his never-dull conversations and company in the graduate school journey. I would also like to thank my undergraduate research assistants Ramadan Ibrahim, Sari Liebhaber, Ileana Kesselman, Rebecca Daniel, and Isha Nagella, for their help with data collection.

Graduate school is where I discovered how much I enjoyed teaching. I need to thank Len Green for his contagious passion for undergraduate education and for enabling graduate students to teach in the department. I also want to thank Julie Bugg and Emily Cohen-Shikora for being the great educators they are and for being role models to look up to. Finally, I want to thank Meg Gregory for all things she taught me on teaching and on being a better educator.

Of course, I need to take a few steps back and acknowledge Ayşecan Boduroğlu and Esra Mungan. They were not only my professors at Boğaziçi University who taught psychology, but they are also the ones who introduced me to research and supported my pursuit of an academic

career. Ayşecan Boduroğlu spent hours with me in her office to help me decide which graduate programs I should apply to. Esra Mungan passionately explained what research was ongoing in the department when I barged into her office one day as a soon-to-be undergraduate student. These are only two small instances that still remind me how caring they are for their students and how enthusiastic they are about what they do. Without their mentorship, I could not have reached the place I am today.

I cannot imagine going through graduate school by myself. I am incredibly thankful to my friends who managed to be supportive and involved in my day-to-day, despite being in different parts of the US and often in different parts of the world. Burcu Yeşil, Aykut Akşit, Anıl Akarsu, Beril Yalçınkaya, Dilşad Özen, Dilan Nebioğlu, Dilşad Sağlam, Cansu Şenocak, and many others—you all are amazing and I could not have survived the past six years without your friendship. I was also lucky to have made excellent friendships during my time in St. Louis. Reshma Gouravajhala, Sam Chung, Eric Failes, Necip Üner, and Marta Stojanovic—though we are already in different cities, I know full well that the distance will not matter. At this point, there are two other people I need to thank. I moved to St. Louis with one of my closest friends, Eylül Tekin; we were labmates and roommates for six years, and parents to the best cats—Tahin and Pekmez—for the last three years of it. I cannot thank Eylül enough for her friendship, and I hope she knows how much she means to me. I also met one of my best friends and my partner in crime during graduate school. The few sentences I am writing here will not be sufficient to thank Francis Anderson. He has been with me through the goods and the bads, and his humor and thoughtfulness are probably what got me through graduate school.

Last, but certainly not the least, I want to thank my parents for being the loving and supportive people they are. From them, I learned how to enjoy life, to work hard toward

something you care about, to build meaningful relationships, and countless other things. They have always been there for me, and I will always strive to be the wonderful people they are.

Oyku Uner

Washington University in St. Louis

August 2021

Dedicated to Elçin and Bülent—aka annişko and babişko.

ABSTRACT OF THE DISSERTATION

Does the Combination of Spacing and Testing Promote Transfer Beyond Either Strategy Alone?

by

Oyku Uner

Department of Psychological & Brain Sciences

Washington University in St. Louis, 2021

Professor Henry L. Roediger, III, Chair

Testing and spacing improve long-term retention and their combination boosts retention further. Despite the combined benefits of spaced testing, it is unclear whether these benefits extend to situations where students learn from lengthy and complex textbooks and need to use concept knowledge in novel ways. To address this issue, in the current study, college students were asked to read from a textbook and review key concepts twice, either back-to-back within the same session or in two sessions spaced two days apart. To review concepts, students either took definition quizzes with feedback (short-answer in Experiment 1, multiple-choice in Experiment 2) or restudied concept definitions. Two days after the last review, students took a short-answer criterial test that included definition and novel application questions. Both quiz formats improved performance on definition questions; however, quizzing benefits were less robust on novel application questions. Multiple-choice quizzes did not improve novel application performance and short-answer quizzes only did so when definition questions preceded application questions on the criterial test. Furthermore, spaced review did not improve performance on either question type. These findings present limitations of retrieval practice and distributed practice as study techniques when students' goal is to flexibly use what they learned in novel ways.

Chapter 1: Introduction

College students often need to study chapters from textbooks to prepare for exams. Consider students in an introductory psychology course who need to learn concepts such as operant conditioning for an upcoming exam. Some exam questions will require students to remember information directly from the textbook, such as the definition of operant conditioning. To successfully answer questions like this, students must have good retention of the relevant textbook chapters. However, some of the questions will require students to use information from the textbook in novel ways, such as explaining how to train a puppy to roll over using the principles of operant conditioning. To successfully answer questions like this, students must apply their concept knowledge to a new case or they must make inferences based on a familiar example. That is, students should be able to transfer what they learned from the textbook to novel contexts. So, how should students prepare for exams that involve both types of questions?

Prior research on study strategies has mostly explored how students should study if their goal is to remember information directly from the learning material and identified several strategies that improve long-term retention (for reviews, see Dunlosky et al., 2013; Fiorella & Mayer, 2016; Miyatsu et al., 2018; Roediger & Pyc, 2012; Pashler et al., 2007). Two of these study strategies in particular are repeatedly shown to have high utility: retrieval practice and distributed practice (Dunlosky et al., 2013). Retrieval practice, also known as testing, refers to the finding that retrieving previously learned information enhances long-term retention, usually relative to restudying the same information (Roediger & Karpicke, 2006; for a review, see Karpicke, 2017). Distributed practice refers to another finding that distributing study of the same material over time enhances long-term retention relative to studying the same material in close

succession (see Cepeda et al., 2006 and Maddox, 2016 for reviews). Because retrieval practice specifies the kind of study and distributed practice specifies its timing, a combination of both strategies promotes better retention than either strategy alone. That is, spacing out practice tests over time improves retention more than massed practice testing or spaced rereading (e.g., Balota et al., 2007; Carpenter & DeLosh, 2005; Carpenter et al., 2009; Cull, 2000; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007). Given the combined benefits of distributed practice and retrieval practice on long-term retention, a widespread recommendation for students is to test themselves on course material repeatedly with repetitions spaced out in time (e.g., Dunlosky et al., 2013; Kang, 2016). The impetus for this dissertation was to explore whether spaced testing benefits extend to situations in which students need to use the learning material to make novel applications and inferences. First, a brief overview of retrieval practice and distributed practice benefits on long-term retention are provided, and then the evidence for the effectiveness of these strategies for transfer of learning are discussed.

1.1 Testing Improves Long-Term Retention

In the past two decades, numerous studies have shown that retrieval of previously learned information improves memory for that information (i.e., *testing effect*, Adesope et al., 2017; Karpicke, 2017; Roediger & Karpicke, 2006; Rowland, 2014). In one such study, after reading two brief passages, college students restudied one of the passages and recalled everything they could from the other (Roediger & Karpicke, 2006). On an immediate test, students who had restudied a passage performed better than those who had recalled the passage; however, on a test two days and one week later, students who had recalled the passage performed better than those who had restudied the passage. This benefit of taking tests on long-term retention is especially striking considering that no feedback was provided to students after initial recall. That is,

although the tested group saw the passages fewer times, they retained more than the restudy group on delayed tests. Findings like this strongly suggest that tests do not simply assess learning; instead, they can be a powerful tool to enhance learning.

Testing as a study strategy has high utility not only because it can substantially boost retention, but also because the testing effect is robust across different situations (Dunlosky et al., 2013; Rowland, 2014). Testing improves long-term retention relative to restudying the learning material (Rowland, 2014) or relative to other study strategies (e.g., concept-mapping, Karpicke & Blunt, 2011, note-taking, McDaniel et al., 2009; see Adesope et al., 2017 for a meta-analysis including control conditions other than restudy). Furthermore, testing enhances long-term retention for ages ranging from preschool (e.g., Fritz et al., 2007) to older adulthood (Tse et al., 2010), though most studies are conducted with college students. In addition, the benefits of testing extend to various learning materials such as word lists (e.g., Carpenter & DeLosh, 2006), paired associates (e.g., Carrier & Pashler, 1992), prose (e.g., Roediger & Karpicke, 2006), as well as pictures (Wheeler & Roediger, 1992), maps (Carpenter & Pashler, 2007), functions (Kang et al., 2011), and visual categories (Jacoby et al., 2010), among others. In sum, retrieval practice improves long-term retention relative to different comparison conditions, for all age groups, and for different learning materials.

An important consideration regarding testing effects is the level of initial test performance. How much participants recall during an initial test is a proxy of their re-exposure to the learning material, particularly in the absence of feedback. Comparing retrieval practice to restudy of the learning material, which typically means restudy of all material or parts of the material relevant to the later test, may create an unfair comparison especially when performance during retrieval practice is low. For example, according to Rowland's meta-analysis (2014),

testing without feedback improved retention compared to restudy only when initial test performance during retrieval practice was greater than 50%.

One way in which initial test performance is manipulated in testing effect studies is employing different initial test formats. For instance, in the absence of feedback, performance on a multiple-choice test is typically greater than performance on a short-answer test, leading to larger testing effects with initial multiple-choice tests (e.g., Kang et al., 2017; Little et al., 2014). However, this pattern can reverse when feedback is provided (e.g., Kang et al., 2007) or differences in testing effects caused by different formats can disappear (e.g., Little et al., 2014). This is related to the consideration that initial test format is often associated with the effort required from participants during retrieval practice. Although most test formats improve retention, tests that require participants to generate an answer (e.g., free recall or short-answer tests) tend to improve retention more than tests that require selecting an answer among alternatives (e.g., recognition or multiple-choice tests) (Carpenter & DeLosh, 2006; Rowland, 2014). However, restudy opportunities in naturalistic settings or the provision of feedback in the laboratory can dictate how different test formats compare to one another in terms of their retention benefits (e.g., Kang et al., 2007; Little et al., 2012; McDaniel et al., 2012; McDermott et al., 2014).

Notably, repeated testing improves retention more than a single test does (Karpicke & Roediger, 2007, 2008; Rawson & Dunlosky, 2011; Rowland, 2014). As an example, in one study, additional tests on the learning material boosted retention by about 150%, whereas additional restudy did not boost retention at all (Karpicke & Roediger, 2008). Findings that show increased benefits of repeated retrieval practice are especially important given that most students prefer to drop out learning material when it has been recalled once (e.g., Karpicke, 2009). So, the

advice for students is to repeatedly practice retrieving the learning material, even if the first retrieval attempt is successful. Using repeated testing, however, leads to a key question regarding its implementation: When should the repetition(s) occur?

1.2 Spacing Improves Long-Term Retention

The *timing* of study can be just as important as the *kind* of study; distributing study opportunities over time enhances long-term retention relative to restudying the same material in closer succession (Cepeda et al., 2006). For example, a meta-analysis that examined 254 studies showed that spacing led to roughly 11% better retention than massing (Cepeda et al., 2006). This retention benefit of spaced study over massed study is referred to as a *spacing effect*. A *lag effect*, on the other hand, refers to the retention benefit when there is greater spacing (i.e., a longer lag) between two or more study opportunities than less spacing (i.e., a shorter lag). In other words, studying the same material with some time passing in between typically improves memory for that material, and furthermore, allowing more time to pass before restudying the material usually improves memory for that material even more. Benefits of distributed practice (both spacing and lag effects) have been investigated since the late 1800s (Ebbinghaus, 1885/1964) and are still frequently explored today (for recent reviews, see Benjamin & Tullis, 2010; Cepeda et al, 2006; Delaney et al., 2010; Maddox, 2016).

Spacing effects are ubiquitous (Cepeda et al., 2006; Dunlosky et al., 2013). Spacing improves retention for all learners across the lifespan; children in preschool (Toppino et al., 1991) up to older adults (Balota et al., 1989) all benefit from distributing practice of learning material over massing. Furthermore, spacing of many types of material boosts retention for that material, including foreign language vocabulary (e.g., Bahrick & Hall, 2005), texts (e.g., Rawson & Kintsch, 2005), pictures (e.g., Hintzman & Rogers, 1973), categories (Birnbaum et al., 2013),

and functions (McDaniel et al., 2013), to name a few. In addition, spacing effects are observed on both immediate and delayed criterial tests (Cepeda et al., 2006). Although some studies show that massing is more beneficial on immediate tests than spacing (e.g., Rawson & Kintsch, 2005), only twelve out of 271 comparisons in a meta-analysis showed null or negative effects of spacing on immediate tests (Cepeda et al., 2006).

When distributing practice, there are several important decisions students need to make: How much time should students allow before reviewing the material? When should students start spacing relative to the exam they are studying for? What kinds of strategies should students be using when they have at least two spaced learning sessions? As it turns out, the answers to the first two questions are related; the optimal spacing gap or lag (i.e., the time between each learning session) is determined by the retention interval (i.e., the time between the last learning session and the criterial test; Cepeda et al., 2008). In a considerable endeavor, Cepeda and colleagues crossed six different lags (ranging from zero to 105 days) with four different retention intervals (ranging from one week to 50 weeks) for learning obscure trivia facts. The optimal lag differed depending on the time of the criterial test, but it fell within 10-20% of the retention interval. In other words, longer lags were needed for longer retention intervals, but criterial test performance declined after increasing the lag more than what was optimal (Cepeda et al., 2008).

1.3 Spaced Testing Improves Long-Term Retention Relative to Spacing or Testing Alone

One of the decisions students need to make is the strategy to use when distributing learning opportunities over time. Critically, some studies have shown that spaced tests promote much better retention than spaced restudy (e.g., Carpenter & DeLosh, 2005; Carpenter et al., 2009; Cull, 2000; Karpicke & Roediger, 2007) or massed (or less spaced) tests (e.g., Bahrick et al.,

1993; Balota et al., 2006; Carpenter & DeLosh, 2005; Cepeda et al., 2008; Cull, 2000; Kapler et al., 2015; Karpicke & Bauernschmidt, 2011). In other words, combining testing and spacing promotes better long-term retention than either strategy alone. As an example, one study showed that learning Swahili-English word pairs with three spaced tests increased delayed cued recall performance (recalling the English word given its Swahili translation) by about 200% relative to three massed tests (Karpicke & Bauernschmidt, 2011). Furthermore, increasing the spacing between each test also increased retention. In another study, middle schoolers who practiced historical facts through spaced tests recalled significantly more on a test nine months later than the middle schoolers who practiced through spaced rereading (Carpenter et al., 2009). Although spaced practice alone (regardless of the kind of practice) can lead to substantial improvements in retention, retrieval during spaced practice is critical: Spaced testing is a more potent tool to improve retention than spaced restudy or massed testing.

The benefits of spaced testing are robust. Distributing retrieval opportunities over time improves memory relative to distributed study or massed (or less spaced) tests for word lists (Karpicke & Roediger, 2007), word pairs (Bahrick et al., 1993; Balota et al., 2006; Cull, 2000; Karpicke & Bauernschmidt, 2011), face-name pairs (Carpenter & DeLosh, 2005), historical facts (Carpenter et al., 2009), content in a natural science course (Kapler et al., 2015) as well as in a college-level math course (Hopkins et al., 2016; Lyle et al., 2019).

Given the combined benefits of testing and spacing on long-term retention, many researchers now recommend students to repeatedly test themselves on course material and to distribute these repetitions over time (Dunlosky et al., 2013; Hopkins et al., 2016; Kang, 2016; Kapler et al., 2015; Lyle et al., 2019; Rawson & Dunlosky, 2011). A large disconnect exists, however, between this recommendation and the ways in which students typically use course

content both inside and outside the classroom. Specifically, most research on retrieval practice and distributed practice has assessed students' verbatim memory of the learning material rather than their deeper understanding of the material. This could be in part due to the simple verbal materials used in past research, such as word lists, word pairs, and isolated statements, which do not allow for an assessment of deeper understanding that could be gained from studying more complex materials. However, one of the key goals of education is to equip students with flexible knowledge so that they can use course material when faced with new problems (Day & Goldstone, 2012; Gick & Holyoak, 1987; Goldwater & Schalk, 2016; Pan & Rickard, 2018). Therefore, determining whether the study strategies that effectively improve memory for text also improve its deeper understanding is critical. Although many researchers recommend students to use distributed practice and retrieval practice for these outcomes, there is not enough evidence to advocate spaced restudy or spaced testing if students' goal is to make novel applications or inferences using information from texts. Put differently, it is unclear whether the retention benefits of distributed practice—and particularly of spaced testing—extend to deeper understanding of lengthy and complex texts.

1.4 The Importance of Transfer in the Classroom

Transfer can be broadly defined as using previously learned information in a novel context (Barnett & Ceci, 2002; Gick & Holyoak, 1987; Perkins & Salomon, 1992). However, this definition does not specify what the learned information is, the context in which it is used, or what aspects of the transfer context are novel. In a seminal review, Barnett and Ceci (2002) captured the breadth of transfer by specifying content (i.e., what is transferred) and context (i.e., where from, where to, and when transfer happens) as two dimensions of transfer, where content and context were further divided into subdimensions (e.g., what performance change is taken as

a measure of transfer, what knowledge domain the content is transferred to). This framework allowed the classification of seemingly different studies as exploring one or several aspects of transfer.

Transfer in the classroom can mean several different things, ranging from students attending a lecture in one room and taking a test in another room (transfer to a different physical context) to students learning the material from their textbook and applying information from the textbook to novel scenarios (transfer to a different knowledge domain). In the current study, the focus was on the kind of transfer that involves flexible use of acquired knowledge on novel application and inference questions, where students need to go beyond verbatim retention of learned material and show mastery of the conceptual content.

On a typical exam, college students may be asked information directly from the course material (e.g., key term definitions from a textbook or practice questions from a study guide) or they may need to use course material in novel ways (e.g., application, inference). As reviewed above, retrieval practice and distributed practice are excellent study strategies to successfully answer the former type of questions. A bigger challenge for students is to use course material beyond the context in which it was originally presented (e.g., Butler et al., 2013; McDaniel et al., 2015; McDaniel et al., 2013).

Assuming students have learned the relevant information at some point prior to a test, several issues can cause transfer failure (Barnett & Ceci, 2002). Students may fail to recognize that the task requires information they already know. For instance, a test question asking how to train a puppy to do a trick requires knowledge about operant conditioning. Students may not immediately recognize what is required, even though they may know what operant conditioning is. Put another way, they cannot access the information in a new context even though it is

available (Tulving & Pearlstone, 1966). Alternatively, students may know that the question targets a specific topic (or they may be provided with this information), but they may fail to recall relevant information about that topic. For example, some students may realize that answering the puppy training question requires using knowledge about operant conditioning, but they may not recall how positive reinforcement works. Furthermore, even if students recall how positive reinforcement works, they may fail to flexibly use their knowledge to provide an appropriate answer. All these cases exemplify what is referred to as the inert knowledge problem, where learners cannot spontaneously and accurately use the information outside the context in which it was learned (Bransford et al., 1989).

In the current study, no distinction was made among the possible reasons for transfer failure listed above. However, it should be noted that the majority of questions assessing transfer provided participants with the relevant concept name within the question itself. As such, the results are more pertinent to the kind of transfer that primarily requires students to remember what a given concept is and then use it to answer novel questions.

Given the importance of deep understanding of course material and the difficulty in achieving it, what are the study strategies that foster transfer? Do study strategies that effectively improve long-term retention also improve transfer? Specifically, do retrieval practice, distributed practice, and a combination of both promote deeper understanding as well as they promote long-term retention? Although testing improves meaningful learning from text (see Pan & Rickard, 2018 for a meta-analytic review), there is not much direct evidence suggesting that distributed practice does. Furthermore, whether combining the two strategies improves meaningful learning beyond either strategy alone has not been established. The goal of this dissertation, therefore,

was to explore whether powerful tools that enhance verbatim retention also enhance transfer to novel application- and inference-type questions after learning from a lengthy and complex text.

1.5 Does Testing Improve Transfer?

Testing is a good study strategy to remember information directly from the learning material; however, research on testing effects has been criticized for not moving beyond verbatim retention (e.g., Butler, 2010; Roediger & Butler, 2011). Until recently, most testing effect studies examined whether taking an initial test improves performance on the same test relative to restudying the learning material. As such, it was unclear whether testing simply improved retention of a specific response or promoted broader retention and understanding of the learning material. Given that instructors do not only assess verbatim retention (Wooldridge et al., 2014) and that deeper understanding of the learning material is an important goal in education (Day & Goldstone, 2012; Goldwater & Schalk, 2016; Pan & Rickard, 2018), it was imperative to evaluate whether testing improves learning outcomes beyond verbatim retention.

Fortunately, many studies in the last decade have shown that testing does enhance transfer to application- and inference-type questions when learning from texts (e.g., Agarwal, 2019; Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, 2016; Hinze & Rapp, 2014; Karpicke & Blunt, 2011; McDaniel et al., 2015; McDaniel et al., 2009; Wooldridge et al., 2014). As one example, participants in one study read prose passages and they were either given repeated tests or were asked to repeatedly study the passages (or repeatedly study isolated facts) before taking a criterial test one week later (Butler, 2010). Repeated testing not only improved performance on previously tested questions, but also on new inference questions from the same domain and from a different knowledge domain. Testing benefits on transfer has been shown not only for learning from text material, but also for learning from videotaped lectures (Butler et al.,

2017), multimedia instruction (Johnson & Mayer, 2009), visual categories (Jacoby et al., 2010, as measured by category induction), and mathematical functions (Kang et al., 2011, as measured by interpolation and extrapolation). Importantly, a recent meta-analytic review showed that testing significantly improves performance on application and inference questions relative to restudy (Cohen's $d = 0.32$); however, testing benefits on verbatim retention are much larger (Cohen's $d = 0.68$, see Pan & Rickard, 2018 for further explication).

Critically, Pan and Rickard's meta-analysis findings (2018) suggested that test-enhanced transfer depends on three factors: level of initial test performance, whether initial test conditions are elaborative, and response congruency across initial and criterial tests. Specifically, higher initial test performance was associated with better transfer. The authors argued that better initial test performance might not only indicate better memory isolated to the tested information, but a more complete memory of the full learning material (which includes the tested information). Pan and Rickard also showed that more elaborative initial test conditions led to better transfer. Though the definition of elaborative conditions is not readily apparent, the authors noted that conditions that include broad or explanatory recall instructions, initial tests that assess higher-order learning (application instead of fact questions), or elaborative feedback (e.g., restudy opportunities after the test, explanation feedback) qualify as elaborative initial tests. Finally, Pan and Rickard also showed that test-enhanced transfer was greater when initial and criterial test responses were identical. However, when defining test-enhanced transfer as tests' influence on application or inference questions, initial and criterial test responses are likely different (but see McDaniel et al., 2015). Therefore, response congruency largely applies to different types of test-enhanced transfer, such as transfer across test formats or transfer to related word cues. Nonetheless, the findings of the meta-analysis suggest that increasing initial test performance

and using elaborative retrieval practice conditions should improve deeper understanding of learning material, as measured by application and inference questions.

1.6 Does Spacing Improve Transfer?

Unlike testing, there is not sufficient evidence to claim that distributed practice improves transfer for learning from text. For instance, Dunlosky and colleagues aptly noted that "although studies using (these) basic measures of memory can inform the field by advancing theory, the effects of distributed practice on these measures will not necessarily generalize to all other educationally relevant measures. Given that students are often expected to go beyond the basic retention of materials, this gap is perhaps the largest and most important to fill for the literature on distributed practice." (2013, p. 38). Even since 2013, the question of whether spacing effects extend to meaningful learning outcomes (e.g., application, inference) has not been fully answered.

To date, only one study investigated distributed practice benefits on text comprehension (Rawson & Kintsch, 2005). In this study, college students read a passage once, read a passage twice back-to-back (i.e., massed rereading), or read a passage twice one week apart (i.e., spaced rereading). Half the students took an immediate comprehension test that required integration, inferencing, and application. The other half took the same test two days later. On the immediate test, students who read the passage twice in a massed fashion performed better than the students who read the passage once and those who read it again in a spaced fashion. However, on the delayed test, students who read the passage twice in a spaced fashion performed better than the students who read the passage once and those who read the passage twice in a massed fashion.

Findings from Rawson and Kintsch (2005) suggest that spaced rereading improves comprehension on a delayed test relative to massed rereading; however, several issues raise concerns for the applicability of these findings. First, participants were given a free recall test

prior to the comprehension test, where they had to write down everything they could remember from a section of the passage. However, testing is not a neutral event; it can improve retention, as well as transfer (Karpicke, 2017; Pan & Rickard, 2018). Therefore, performance on the comprehension test might have been influenced by the recall test that occurred prior rather than the spacing that occurred between readings of the passage. Specifically, considering that spaced rereading indeed led to better performance on the delayed recall test than massed rereading, better comprehension after spaced rereading could simply be due to the better recall observed on the preceding test. Furthermore, the kind of rereading participants did in this study may not be representative of how students would reread lengthier texts. Though rereading entire passages is feasible with the 1,730-word text used in this study, it would be impractical for students to reread entire chapters from a textbook. Instead, students may choose to reread parts of lengthy texts. Finally, no study has replicated the key finding from this study that spaced rereading improves comprehension performance at a delay.

Importantly, whether distributed practice improves text comprehension assumes that reviewing a text is necessary to boost comprehension. Critically, though, whether rereading enhances comprehension from texts is unclear (Miyatsu et al., 2018), and the complexity of the text may matter (Callender & McDaniel, 2009). Some studies have shown improvements on comprehension tests due to rereading (Barnett & Seefeldt, 1989; Karpicke & Blunt, 2011; Rawson et al., 2000; Rawson & Kintsch, 2005), whereas other studies did not find rereading benefits (e.g., Agarwal, 2019; Callender & McDaniel, 2009; Griffin et al., 2008). As an example, Callender and McDaniel had students study several different texts (including the one used in Rawson and Kintsch's study) in massed conditions across four experiments, and they did not obtain rereading benefits on comprehension on an immediate or a delayed test. Based on their

findings, they concluded that rereading benefits may not extend to the kinds of texts students need to learn from or to the kinds of tests they are assessed on (Callender & McDaniel, 2009). However, rereading in all these studies occurred in one schedule (typically in a massed fashion); that is, spacing was not manipulated (but see Rawson & Kintsch, 2005). Therefore, *when* rereading takes place may be a critical factor in understanding its effects on comprehension.

Despite the scarce evidence for the benefits of distributed practice on transfer of learning from texts, studies exploring spacing or lag effects with different materials or tasks suggest that distributed practice may in fact improve deeper understanding. As an example, category learning studies typically show that spaced rather than massed presentation of category exemplars leads to better category induction, which is a measure of transfer in this domain (Carvalho & Goldstone, 2015, 2019; Kornell & Bjork, 2008, but see Carpenter & Mueller, 2013; Carvalho & Goldstone, 2014; Zulkiply & Burt, 2013). In a typical category learning experiment, participants study multiple exemplars from at least two categories. To measure category learning, participants are then asked to categorize previously presented exemplars (memory items) and novel exemplars from studied categories (transfer items). This process is akin to students classifying a new example as a concept they have learned before (e.g., training a puppy to do a trick is an example of operant conditioning) and is therefore an essential part of learning and instruction.

Though the spacing benefit in category induction indirectly suggests that spacing may improve transfer of learning from text, spacing in category learning experiments is often confounded with interleaving; that is, exemplars from other categories are presented in between exemplars from the same category to achieve spacing. In fact, several studies have shown that it is the category alternation that improves transfer rather than temporal spacing between exemplars from the same category (e.g., Birnbaum et al., 2013; Kang & Pashler, 2012, but see Foster et al.,

2019). Fortunately, though, some category learning experiments do show that increased temporal spacing, when keeping category alternations constant, can also enhance transfer (Birnbaum et al., 2013; Carvalho & Goldstone 2015). In one such study, participants were presented with exemplars from different butterfly species in an interleaved fashion (Birnbaum et al., 2013). Some participants were presented with three exemplars from other species in between exemplars from the same species (i.e., short lag), whereas other participants were presented with fifteen (i.e., long lag). The condition with the longer lag led to better category induction, suggesting that distributed practice can have a unique contribution to transfer in the context of category learning. However, distributed practice in category learning experiments does not involve the repetition of the same item (as is the case in research with verbal materials), but a new instance of a to-be-learned category.

Similar to category learning, function learning also benefits from distributed practice (McDaniel et al., 2013). In function learning tasks, participants learn input-output pairs (e.g., [0.5, 2] and [2, 8] if the function is $y = 4x$), typically through feedback training (i.e., they guess the output given the input, followed by feedback; Busemeyer et al., 1997). Participants are then tested on input-output pairs they have previously seen (memory items) and on new pairs that fall within the function (transfer items). As one study showed, when participants were trained on a function using spaced practice (inputs were repeated only after all inputs were presented once) rather than massed practice (only one input intervened another input and its repetition), they were better able to provide outputs to new input values both within and outside the training range (McDaniel et al., 2013). Though the same items are repeated in function learning tasks (unlike category learning experiments), functions are certainly much different than texts as learning materials.

Although distributed practice improves transfer in domains such as category and function learning (see also spacing effects in math learning, e.g., Rohrer & Taylor, 2006, 2007), the materials used in these studies differ from the lengthy and complex texts (i.e., textbook chapters) students often need to study. An open question, therefore, is whether distributed practice effects on transfer extend to text materials.

1.7 Does Spaced Testing Improve Transfer Relative to Spacing or Testing Alone?

Distributed practice and retrieval practice are two powerful study strategies (Cepeda et al., 2006; Dunlosky et al., 2013; Rowland, 2014) and researchers advocate combining the two for even better learning outcomes, based on findings that spaced testing improves long-term retention relative to either strategy alone (Bahrick et al., 1993; Balota et al., 2006; Carpenter & DeLosh, 2005; Carpenter et al., 2009; Cepeda et al., 2008; Cull, 2000; Kapler et al., 2015; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007). However, as described in earlier sections, the studies that show the combined benefits of spacing and testing typically use learning materials less complex than what students may need to study, and they only assess verbatim retention rather than comprehension. Given the mismatch between the kinds of criterial tests used in prior research and the kinds of assessments students receive in the classroom (as well as how they would use course material outside the classroom), whether spaced testing enhances students' ability to make novel applications and inferences from lengthy and complex texts remains an important issue to be addressed.

Although some previous studies have employed spaced testing and examined transfer performance, these studies were not designed to assess whether the combination of spacing and testing promotes transfer beyond either strategy alone. For instance, one study examined the

effectiveness of an intervention implemented on homework assignments within a college engineering course (Butler et al., 2014). The intervention required students to complete three times as many practice problems relative to the usual homework assignments, and the problem sets were distributed across weeks. Feedback was available to students immediately after the homework deadline and students were required to look at the feedback to get points. The intervention was compared to standard practice for homework assignments in the course, where there were fewer practice problems, no review of the material in the following weeks, no immediately available feedback, and no requirement to look at the feedback. Butler and colleagues found that the intervention significantly improved performance relative to standard practice on two take-home exams, which included both free-form and multiple-choice questions assessing novel application of course concepts.

This study suggested that more tests and spaced tests can improve transfer relative to fewer tests and massed tests; however, it did not tease apart the possible unique contributions of testing and spacing. In designs like this, where spaced testing conditions are compared to no spacing or testing, only spaced restudy, or only massed testing, it is not possible to discern whether the combination of the two strategies enhances transfer more than either strategy alone.

1.8 Theoretical Considerations

Many theories have been proposed to explain retrieval practice and distributed practice effects. A survey of these theories, as well as evidence for or against them, is not within the scope of this dissertation; however, two issues are relevant. First, the mechanisms used to explain the benefits of testing and spacing have some overlap, leading to the possibility of redundancy of these strategies when measuring transfer of learning. Second, existing theories of testing and spacing

rarely make predictions regarding the transfer of learning as a separate outcome from the verbatim retention of previously learned material.

Of the current theories of spacing, study-phase retrieval (Benjamin & Tullis, 2010), encoding variability, and in particular their combination, is considered to account for many of the findings in the literature (Delaney et al., 2010; Maddox, 2016). According to the study-phase retrieval account, a second presentation of an item reminds learners of its first presentation (or prior presentations of other related items) and triggers its retrieval. Critically, if items are repeated back-to-back, learners will recognize the repetition from immediate memory; however, if items are spaced apart in time, learners will retrieve a prior presentation from long-term memory. Given the effort involved in retrieving from long-term memory, learners will benefit from spaced (rather than massed) presentation to the extent that an item repetition reminds learners of a prior presentation of the item. Thus, the study-phase retrieval account suggests that spacing is effective relative to massing (or less spacing) because participants engage in covert retrieval during the repetition of an item. Put differently, according to this account, there is overlap in the mechanism through which testing and spacing effects emerge. Importantly, if the gap between repetitions is too large (see Cepeda et al., 2008, for discussion of optimal spacing gaps), this account predicts that learners will be less likely to be reminded of an earlier presentation, and this in turn will eliminate any retention benefits of spacing.

Another mechanism considered to underlie spacing effects is the variability in contextual factors when items are repeated with a larger gap in between, relative to when they are presented in close succession. The encoding (or contextual) variability account suggests that increasing the contextual factors with which an item is encoded provides the learner with more routes to retrieve it later. As context is considered to change over time, larger gaps between repetitions

should lead to more variable encoding, and therefore better memory for items encoded in these variable contexts. However, spacing effect studies have consistently demonstrated that an optimal spacing gap exists for a given retention interval. That is, increasing the gap between repetitions improves retention until a certain point, and retention is impaired with increasing the gap beyond that point (i.e., an inverted U-shape for the optimal spacing gap)—a finding that challenges the encoding variability account. In fact, Appleton-Knapp et al. (2005) found that varying the context improved memory when the spacing gap was short, whereas it impaired memory when the spacing gap was long, relative to keeping the context the same (in this study, context was the layout on which an item was presented).

Nonetheless, given the strengths of both the study-phase retrieval and the encoding variability accounts, researchers have also proposed that a combination of both accounts explains the majority of the findings in the spacing effect literature (Delaney et al., 2010; Raaijmakers, 2003; Siegel & Kahana, 2014). According to these multi-process accounts, to the extent that an item repetition can remind learners of its prior presentation, the increased contextual variability through longer gaps between repetitions should improve memory. That is, spacing has mnemonic benefits over massing (or less spacing) because of the benefits of retrieval on memory and the increased retrieval routes afforded by variable encoding contexts.

The similarities among theories explaining testing and spacing effects become particularly salient when considering one of the recent theories of testing effects: the episodic context account (Karpicke et al., 2014). Despite the prominence of context-based accounts of spacing, Karpicke et al. pointed out that no account of testing refers to the role of contextual factors. The episodic context account assumes that learners encode temporal or episodic context information along with to-be-learned items, and that during retrieval, learners reinstate the

context in which items were initially learned. Also during retrieval, learners' memory for items is assumed to be updated to incorporate contextual information from both the prior presentation being retrieved as well as the retrieval context. Finally, the varied contextual information associated with an item is assumed to help in its retrieval at a later point (similar to the encoding variability account of spacing). The episodic context account argues that retrieval allows learners to integrate contextual information from different presentations (i.e., from the initial study phase and the retrieval practice phase) and because these contexts differ, the variability in context promotes better subsequent memory.

The episodic context account of testing effects and the multi-process accounts of spacing effects (which combine encoding variability and study-phase retrieval) share many similarities. Both make the assumptions that context changes over time, that retrieval of a previously presented item during its repetition leads to encoding of additional contextual information, and that this variability in the encoded contextual information benefits memory later. However, Karpicke et al. (2014) noted that the episodic context account differs from encoding/contextual variability and study-phase retrieval accounts of spacing effects in one critical aspect: Retrieval in testing effect experiments is *intentional,* whereas retrieval of the study phase during spaced repetitions of items is *incidental*. Accordingly, Karpicke et al. argued that this intentional retrieval should lead to a greater contextual updating than the incidental retrieval triggered by spacing. Nonetheless, the similarities among these accounts point to a possible shared mechanism of these effects.

Aside from the similarity of some theories explaining testing and spacing effects, another important consideration is that these theories rarely make predictions regarding the transfer of learning. Even if what is meant by transfer is constrained to instances in which learners must use

their knowledge to answer novel application and inference questions (as in the current study), theories of spacing and testing rarely differentiate this learning outcome from the retention of information as it was initially learned. This is to be expected to some extent for theories of spacing effects or lag effects, given that most prior studies have used simple materials such as word pairs, where novel application and inference could not be assessed.

Within the testing effect literature, prose passages are frequently used as learning material, and especially more recently, researchers have examined whether the benefits of testing extend to previously studied but non-tested information, as well as whether learners can use previously studied information in novel ways on application and inference questions. Despite the increasing number of testing effects studies on transfer, however, theories of testing have been slower to emerge, and existing theories do not differentiate between retention of previously studied information and the application of that knowledge to new contexts.

Of the available explanations, elaboration-based mechanisms, where tests instigate a semantic elaboration process that creates more retrieval routes for later (Carpenter, 2009), can potentially explain why testing might improve performance on novel application and inference questions. To the extent that these elaborations include or are similar enough to an application question's answer, learners should be better equipped to transfer their knowledge after an initial test than after restudying. However, knowledge of a concept can be assessed in a variety of ways, and therefore it seems unlikely that learners' elaborations will always aid in novel application and inference questions. Hence, elaboration-based accounts cannot systematically explain when or why tests enhance application of knowledge to novel contexts.

A different approach to the question of whether spaced testing will improve transfer of learning is to consider theoretical frameworks of transfer. According to Barnett and Ceci's

(2002) three-part framework, for example, for successful transfer to occur, learners must *recognize* the relevance of previously learned information given a novel context, *recall* this relevant information, and successfully *execute* or *apply* it in the novel context. Although the recognition step is likely the most important but elusive step (Gick & Holyoak, 1980), participants in testing or spacing research are typically told a criterial test will assess their memory or understanding of material previously studied within the experiment. Thus, the first step of Barnett and Ceci's framework is bypassed in most relevant prior research (though the specificity of cues participants receive might vary), as well as in the current study.

Within Barnett and Ceci's framework, the *recall* step is arguably where spacing and testing show their effects. Given robust benefits of both techniques on long-term retention, spacing and testing should facilitate the transfer process by improving accessibility of the learning material on a later test. However, increased accessibility of previously learned and relevant information may be necessary but not always sufficient for learners to flexibly use the information as required in the *execution* step. Because research on testing and spacing does not isolate the *recall* and *execution* steps, it is difficult to investigate whether these techniques facilitate transfer by improving recall, execution, or both.

Although the extant research suggests that spacing and testing improve the *recall* step, it is still possible that spacing or testing improves the *execution* step as well. According to Butler et al. (2017), however, who used a two-phase criterial test where the second phase eliminated the need to recall information and simply required participants to apply it, benefits of repeated retrieval over repeated studying disappeared. These findings tentatively suggest that testing facilitates transfer due to improved retention of knowledge. In the current study, participants did

not receive a two-phase criterial test; therefore, the *recall* and *execution* steps could not be isolated.

In sum, current theories of testing and spacing are not informative to make predictions regarding whether these techniques will improve transfer of learning to novel application and inference questions, and whether their combination will lead to interactive effects. Currently, these theories suggest that testing and spacing will increase accessibility of the learned material on a later test. It is unclear, however, whether this alone is enough for successful transfer and whether these techniques can separately facilitate the application of retrieved knowledge. Further, the similarity in the proposed mechanisms for testing and spacing effects suggest that combining both (i.e., spaced testing) may not have interactive benefits when assessing transfer of learning.

1.9 Introduction to the Experiments

The primary goal of this dissertation was to examine whether a combination of spacing and testing with feedback improves novel application when studying from a textbook, beyond either strategy alone. Across two experiments, college students first read a textbook chapter, and reviewed key concepts from the chapter twice in one of four ways: massed restudy, massed testing, spaced restudy, or spaced testing. When review was massed, students revised concepts within the same learning session in close succession, whereas when review was spaced, students revised concepts two days apart. To review concepts, students either took quizzes that asked definition-based questions on key concepts and then received correct-answer feedback, or they restudied the definitions of those concepts. Two days after the final review, participants completed criterial tests, where they answered definition questions whose answers they had

initially studied, and application questions that required using conceptual knowledge in novel ways that were not presented in the textbook.

By manipulating review type (restudy or test) and review timing (massed or spaced), and by assessing two different learning outcomes (retention and transfer), this dissertation explored whether strategies that have been shown to effectively improve verbatim retention also improve novel application. Furthermore, by changing initial quiz format from short-answer questions in Experiment 1 to multiple-choice questions in Experiment 2, this dissertation also explored whether differences in initial test format (and thus in initial test performance) influence the combined benefits of spacing and testing on both verbatim retention and novel application.

Based on previous research, on the definition questions within the criterial test, students who took quizzes should outperform those who reread concept definitions (i.e., testing effects), and students should perform better on concepts they reviewed over the course of two days rather than those they reviewed back-to-back within the same session (i.e., spacing effects). Furthermore, based on previous research demonstrating the combined benefits of testing and spacing, students should perform best on the criterial test's definition questions after reviewing concepts via spaced tests.

The critical question is whether the same pattern observed for definition questions on the criterial test will also be observed on application questions. Based on available evidence, students who took quizzes should perform better on application questions than those who reread concept definitions (i.e., testing effects). However, the benefit of testing is expected to be smaller for application questions than definition questions (Pan & Rickard, 2018). It is less clear whether spaced review of concepts will lead to better application question performance than massed review of concepts. Although spacing effects are robust, there are no prior studies that

investigate spacing effects using lengthy and complex texts and that manipulate spacing within a timeline similar to that employed in the current study. It is reasonable to predict, however, that spacing *will* improve novel application when studying text material (see Rawson & Kintsch, 2005, who showed that spaced rereading improves comprehension of brief texts), given that spacing improves novel classification in category learning experiments—a related form of transfer. However, learning from complex texts may involve different processes than learning from simpler materials (Callender & McDaniel, 2009); if so, spaced review may not benefit novel application.

When specifically considering tested participants, relative benefits of spaced and massed review may depend on initial quiz performance (for both definition and application questions). Performance on the second quiz will likely be greater in massed rather than spaced conditions, especially when students receive feedback on the quizzes (see Carpenter & DeLosh, 2005; Carpenter et al., 2009, for similar observations). Because greater initial test performance is associated with greater test-enhanced transfer of learning (Pan & Rickard, 2018), it is possible for massed testing to lead to better application performance than spaced testing. However, correct-answer feedback provided on the last quiz may counteract this effect. Alternatively, novel application after massed and spaced tests may be comparable to one another (i.e., no spacing or massing effects). Although improving retention of previously learned information is important to achieve successful transfer, it is also essential to improve the execution of that recalled information (Barnett & Ceci, 2002). Given that testing and spacing both primarily improve retention, and possibly do so using a shared mechanism, spaced tests may not improve novel application above and beyond massed tests (or spaced restudy). That is, having good memory for a concept's definition may not be sufficient to answer novel application questions.

Chapter 2: Experiment 1

2.1 Method

2.1.1 Participants

Sample size was determined based on a power analysis for a main effect of a between-groups factor in a mixed design, setting power at 0.80 and alpha at 0.05, and assuming a medium effect size for a testing effect on novel application (G*Power 3.1.9.4, Faul et al., 2007). This power analysis called for 98 participants. To have an equal number of participants across the counterbalancing conditions, the sample size was set to 112 participants. 123 Washington University undergraduates (M_{age} = 19.85, SD_{age}= 1.49) completed all three sessions for course credit (2.5 credits) or payment ($20). To incentivize participants to return to follow-up sessions, a bonus payment of $5 was awarded if participants completed all three sessions. Due to the COVID-19 pandemic, in-person data collection was halted at the end of February 2020 and resumed in an online format in March 2020. Of the participants who completed all sessions, 49 participated in the laboratory and 74 participated online[1]. 58 additional participants started the experiment (11 in the laboratory and 47 online); 13 did not finish Session 1, 17 did not return for Session 2, 17 did not return for Session 3, and 11 participants' sessions had to be cancelled due to a programming error.

2.1.2 Design

A 2 x 2 x 2 mixed-factorial design was used, where *review type* (restudy vs. test, between-groups), *review timing* (massed vs. spaced, within-groups), and *question type* (verbatim definition vs. application, within-groups) were manipulated. After reading a chapter from a research methods textbook, half of the participants reviewed concepts through quizzes with

[1] Participants who completed the experiment in the laboratory and those who completed it online did not have any performance differences on the criterial test, suggesting the samples were comparable.

immediate feedback, and the other half reviewed concepts through restudy of concept definitions (these were answers to quiz questions). Participants reviewed half of the concepts in a massed fashion, where they either took a quiz twice back-to-back with feedback (massed testing) or they restudied concept definitions back-to-back (massed restudy). Participants reviewed the remaining half of the concepts in a spaced fashion, where they re-took the quiz two days later with feedback (spaced testing) or restudied concept definitions for the second time two days later (spaced restudy). Concepts assigned to massed or spaced review were counterbalanced such that each concept was reviewed massed or spaced an equal number of times across participants. All participants took a criterial test two days after the second round of review. Critically, concepts were tested with both verbatim definition questions and novel application questions, but the order was counterbalanced so that each question type made up the first half of the criterial test an equal number of times across participants.

2.1.3 Materials

Participants read a 38-page text on research methods. The text was compiled from *Research Methods in Psychology, 3rd edition* (Heiman, 2002) and was borrowed and adapted from two prior studies using these materials (Anderson & McDaniel, 2021; McDaniel et al., 2015). 20 concepts from the textbook were identified and split into two sets of ten concepts for counterbalancing purposes (see Appendix A). Half of the participants reviewed concepts in Set 1 spaced and those in Set 2 massed, whereas the other half reviewed concepts in Set 2 spaced and those in Set 1 massed. In addition, two different orders were randomly generated for the concepts to appear within each set. This was done to ensure that participants reviewed the concepts in the same order across the two review opportunities. Concept order was counterbalanced so that half of the participants reviewed concepts in Order 1 and the other half reviewed them in Order 2.

Finally, one definition and one application question for each concept were created in short-answer format, totaling 40 questions. On criterial tests, participants answered both definition and application questions. Critically, the order of question type on the criterial tests was counterbalanced such that half of the participants first answered definition questions before moving onto application questions, and others first answered application questions before moving onto definition questions. This third counterbalancing was implemented to account for the possible effects of answering one type of question on the other. Definition and application questions corresponding to each of the 20 concepts can be found in Appendix B.

Participants assigned to the Test group answered short-answer definition questions on their quizzes, and they were given correct-answer feedback in the form of concept definitions. Participants assigned to the Restudy group, on the other hand, simply read concept definitions as part of their review.

2.1.4 Procedure

The experiment consisted of three sessions that were each two days apart (e.g., Session 1 on Monday, Session 2 on Wednesday, Session 3 on Friday). As stated earlier, the experimental procedure had to change from being in the laboratory to being online in the middle of data collection due to the pandemic. In-person participants came to the laboratory at timeslots scheduled 48 hours apart and an experimenter was present during all three sessions. Online participants, on the other hand, were provided with the link for Session 1 and completed it at a time of their choosing without an experimenter present. These participants were sent an email with the Session 2 link two days after they completed Session 1, and an email with the Session 3 link two days after they completed Session 2. Participants were asked to complete these follow-up sessions within the same day they were sent the email and, if possible, at the time they started

the previous session(s). However, because there was no experimenter present for the online participants, the three sessions were not always exactly 48 hours apart—a point discussed later. In-person and online data collection procedures were otherwise identical, and any differences between the two are described below where necessary.

Session 1. At the beginning of Session 1, in-person participants were asked whether they had previously taken a research methods class or if they were taking any at the time, and participants who answered yes were not eligible to participate. This screening was not possible for the online participants (the experiment link was posted on the subject pool website, so participants' course history could not be monitored); however, all participants (in-person and online) answered a post-experimental questionnaire asking about possible research methods courses taken, among other questions.

After providing consent, participants were asked whether they were willing to share their SAT or ACT scores. If they agreed to do so, students who participated in the experiment in the laboratory were asked to sign a data release form and students who participated in the experiment online were asked to self-report their scores. SAT and ACT scores were collected to examine whether there was differential attrition between the Test and Restudy groups and to examine if participants who completed all three sessions were different from those that did not.

All participants first read the research methods chapter and they were instructed to simply read the chapter rather than study it in detail. Participants were given one hour to read, but they were allowed to move on to the next phase of the experiment after 45 minutes if they finished reading[2]. Participants were able to scroll through the chapter rather than seeing only one page at a time. After participants completed reading (or after one hour passed), they completed a 5-min

[2] The time given to participants to read the textbook was based on the two prior studies using the same learning material.

filler task. The filler task asked participants to indicate whether 60 names, presented one at a time on the screen for 5-sec each, were US presidents.

Participants then reviewed information from the chapter either through a quiz or restudy, depending on their randomly assigned condition. First, participants reviewed ten of the concepts (either Set 1 or Set 2, and in either Order 1 or Order 2) once. The Test group answered ten short-answer definition questions, where questions appeared on the screen one at a time and participants were required to write a response before moving on to the next question. After each question, participants received correct-answer feedback whether or not they answered the question correctly. The feedback was presented for at least 5-sec, and the program advanced to the next trial after 30-sec elapsed. After participants answered these questions, they were given a new set of definition questions on the remaining set of concepts. As with the first set of questions, participants were forced to respond before moving to the next question and they were presented with correct-answer feedback after each question. However, unlike the first set of questions, participants were given the second question set twice (i.e., massed testing), again with correct-answer feedback provided after each question. The order of questions within each set was the same when presented for the second time. Therefore, the Test group answered 30 definition questions in total, where one set was tested once (i.e., ten questions) and then the other set was tested twice (i.e., twenty questions).

The Restudy group, on the other hand, read the definitions of concepts instead of trying to recall them from memory. All definitions were presented one at a time on the screen for at least 10-sec, and the program advanced to the next trial after 30-sec elapsed. Similar to the Test group, participants first read definitions of concepts from one set once (i.e., ten definitions), and then read definitions of concepts from the other set twice (i.e., twenty definitions; massed

restudy). For the set that was repeated, participants were presented with the definitions in the same order, similar to the Test group.

After reviewing concepts, participants completed a post-experimental questionnaire. They rated their prior knowledge of the contents of the chapter on a 10-point scale and indicated whether they had enough time to read the chapter. For the online participants, these two questions, along with an attention check question, were asked immediately after participants read the chapter[3]. All participants were asked, at the end of Session 1, whether they had previously taken a research methods course or if they were taking one at the time, whether they participated in a study using the same learning material before, and whether they were doing anything else during the experiment. Online participants were asked additional questions pertaining to the environment in which they completed the experiment (e.g., whether they were alone, whether there was music in the background) and to their level of distraction (e.g., whether they left the experiment longer than 5-min, whether they were actively on their phone or computer). The online participants were also asked whether they took notes while reading the textbook or if they recorded it, and whether they thought they gave as much effort or attention to the experiment as they would if they could be in the laboratory.

Session 2. In-person participants returned to the laboratory 48 hours after Session 1, and online participants were sent an email with the experiment link two days after they completed Session 1.

In Session 2, all participants first took a criterial test consisting of ten definition questions and ten application questions on concepts that were reviewed twice in Session 1. Half of the participants first answered all the definition questions before moving onto the application

[3] This change was implemented, because it would not necessarily impact any of the critical components of the design, but in hindsight seemed a better way to assess this information.

questions, and the other half first answered all the application questions before moving onto the definition questions. Feedback was not provided on the criterial test.

After the criterial test, participants reviewed concepts that they only reviewed once during Session 1. Participants either took the same quiz for these concepts with correct-answer feedback (i.e., spaced testing) or they restudied concept definitions (i.e., spaced restudy). Whether participants were in the Test or Restudy group was consistent across sessions, such that those who received a quiz in Session 1 also reviewed concepts with a quiz in Session 2 (and vice versa). In addition, the order of concepts during the review in this session was the same as the order in which these concepts were reviewed in Session 1. The timing of questions, feedback, and definitions were identical to Session 1. At the end of Session 2, online participants were asked additional questions pertaining to the environment in which they completed the experiment and to their level of distraction.

Session 3. In-person participants returned to the laboratory 48 hours after Session 2, and online participants were sent an email with the experiment link two days after they completed Session 2. In this session, all participants took a criterial test on concepts that were reviewed across the two previous sessions. Similar to the criterial test in Session 2, this test consisted of ten definition questions and ten application questions, and feedback was not provided on these questions. Half of the participants first answered definition questions before application questions, and the other half first answered application questions before definition questions. This was consistent across Sessions 2 and 3, such that participants who first answered definition questions on the criterial test in Session 2 also first answered definition questions on the criterial test in Session 3 (and vice versa).

After the criterial test, participants completed a post-experimental questionnaire, where they were asked about any prior research methods courses taken, whether they participated in an experiment using the same learning materials before, whether they were doing anything else during the experiment, and whether they studied the learning material in between sessions. Online participants were asked additional questions, such as whether they got help with answering questions during the experiment, whether they thought they gave as much effort or attention to the experiment as they would if they could be in the laboratory, and other questions about the environment in which they completed the experiment (e.g., whether they were alone, whether there was music in the background) and their level of distraction (e.g., whether they left the experiment longer than 5-min, whether they were actively on their phone or computer).

2.1.5 Scoring

Short-answer responses on quizzes and criterial tests were hand-scored, where correct answers were given 1 point, partially correct answers were given 0.5 points, and incorrect answers were given 0 points based on a rubric. Two raters who were blind to conditions scored approximately 10% of all criterial test responses. The raters showed good inter-rater reliability (weighted Cohen's kappa = 0.87), thus, the remaining questions were scored by one rater.

2.2 Results

123 participants completed all three sessions of the experiment (n = 61 for the Restudy group, n = 62 for the Test group). One participant in the Restudy group who had 0 on the criterial test (they responded "don't remember" on all questions) was excluded from the analyses. Two participants (one in the Restudy and one in the Test group) indicated taking notes while reading the chapter; however, excluding these participants did not change the findings. As such, they are included in the analyses. Fifteen participants (six in the Restudy and nine in the Test group)

indicated that they previously took or that they were currently taking statistics or research methods courses. The analyses include these participants as well, as excluding them did not change any of the findings.

To determine whether participants who did not come back for Sessions 2 or 3 differed from those who completed the experiment, participants' ACT scores were examined. SAT scores were converted to ACT scores, as very few participants had SAT scores alone. Of the participants who had ACT information (range: 24-36, M = 33.38, SD = 1.90), there were no ACT differences between participants who completed the experiment and those who did not: (1) participants who missed Session 2 (n = 12, M = 33.50, SD = 1.38) were no different than those that returned for it (n = 127, M = 33.33, SD = 1.97), (2) participants who missed Session 3 (n = 16, M = 33, SD = 2.42) were no different than those who completed all sessions (n = 111, M = 33.38, SD = 1.90), and (3) Restudy and Test participants who did not return for either session were also similar to one another in terms of ACT scores.

Below, I report results from the quiz and the criterial test using an ANOVA framework. All omnibus tests of statistical significance use an alpha level of 0.05 and all pairwise comparisons are reported with a Bonferroni correction. Effect sizes are reported using partial eta-squared for main effects and interactions, and Cohen's d for pairwise comparisons. To better account for within-person correlations and examine trial-level accuracy, I also report a complementary analysis that used multi-level logistic regression to predict criterial test accuracy. For ease of interpretation, I focus my results on the ANOVAs; to avoid redundancy, I only detail patterns which differed notably from those obtained using the ANOVAs (full model results are reported in tables).

2.2.1 ANOVA Results

Quiz Performance. The Test group took two short-answer quizzes with feedback during Sessions 1 and 2. In Session 1, participants completed one of these quizzes twice back-to-back, in a massed fashion. Participants also completed the other quiz twice, the first time in Session 1 and the second time in Session 2, in a spaced fashion. One participant's quiz data from Session 1 were not recorded due to a programming error. A 2 (quiz number: quiz 1 or quiz 2) x 2 (review timing: massed or spaced) x 2 (which concept set was spaced: set 1 or set 2) x 2 (question order: order 1 or order 2) mixed factorial ANOVA was conducted to examine quiz performance.

There was a marginal main effect of the question order counterbalancing, $F(1, 57) = 3.40$, $p = 0.07$, $\eta_p^2 = 0.06$, where participants performed better on quizzes when questions were in Order 2 than in Order 1. However, because the two orders were created randomly and this variable did not interact with any of the main variables or the other counterbalancing variable, this effect is not discussed further. Which set of concepts was spaced or massed did not have a main effect ($p = 0.31$) and this counterbalancing variable only interacted with quiz number and review timing, which is described below.

Quiz performance based on quiz number and review timing is displayed in Table 1. Overall, participants performed better on Quiz 2 ($M = 0.46$, $SD = 0.17$) than they did on Quiz 1 ($M = 0.35$, $SD = 0.14$), $F(1, 57) = 57.49$, $p < 0.001$, $\eta_p^2 = 0.50$. Further, participants showed similar performance on massed ($M = 0.42$, $SD = 0.18$) and spaced ($M = 0.39$, $SD = 0.15$) concepts ($F(1, 57) = 2.13$, $p = 0.15$, $\eta_p^2 = 0.04$). Critically, there was a significant interaction between review timing and quiz number, $F(1, 57) = 50.65$, $p < 0.001$, $\eta_p^2 = 0.47$. Although participants performed better on spaced concepts ($M = 0.38$, $SD = 0.16$) than they did on massed concepts ($M = 0.33$, $SD = 0.18$) in Quiz 1 ($p = 0.01$, $d = 0.33$), participants performed better on

massed concepts (M = 0.52, SD = 0.21) than they did on spaced concepts (M = 0.40, SD = 0.17) in Quiz 2 ($p < 0.001$, d = 0.69). That is, for massed concepts, performance improved by 19% from Quiz 1 to Quiz 2 (with correct-answer feedback), whereas performance only increased by 2% for spaced concepts.

Table 1 Performance of the Test group on Quiz 1 and Quiz 2 on massed and spaced concepts in Experiment 1.

	Quiz 1	Quiz 2
Massed	.33 *(.18)*	.52 *(.21)*
Spaced	.38 *(.16)*	.40 *(.17)*

Note. Standard deviations are reported in parentheses.

These patterns, however, were qualified by a three-way interaction between which concept set was spaced or massed, quiz number, and review timing, $F(1, 57) = 5.13$, $p = 0.03$, $\eta_p^2 = 0.08$. A closer look at quiz performance indicated that massed concepts were better answered than spaced concepts on Quiz 2 regardless of the concept set that was spaced (both ps < 0.01). However, on Quiz 1, spaced concepts were better answered than massed concepts only when Set 1 was spaced ($p = 0.002$), and there were no differences between spaced and massed concepts on Quiz 1 when Set 2 was spaced ($p = 0.85$). It is unclear why spaced concepts were better answered than massed concepts on Quiz 1 when only Set 1 was spaced, as no differences were expected between massed and spaced items at this point in the experiment.

Criterial Test Performance. Both the Restudy and Test groups took one criterial test on massed concepts during Session 2, and another criterial test on spaced concepts during Session 3. Criterial test performance was examined based on review type (whether participants restudied concepts or were quizzed on them), review timing (whether concept review was massed or spaced), and question type (whether participants answered definition or application questions). The two counterbalancing variables pertaining to the criterial test (whether participants saw

definition or application questions first, whether concepts in Set 1 or Set 2 were spaced) were also included in the ANOVA as between-subjects factors.

A 2 (review type: restudy or test) x 2 (review timing: massed or spaced) x 2 (question type: definition or application) x 2 (which concept set was spaced: set 1 or set 2) x 2 (question order: definition first or application first) mixed factorial ANOVA was conducted to examine criterial test performance. Neither of the counterbalancing variables had a significant main effect (both Fs < 1). However, these variables had some reliable interactions with the main variables, and therefore were not taken out of the ANOVA.

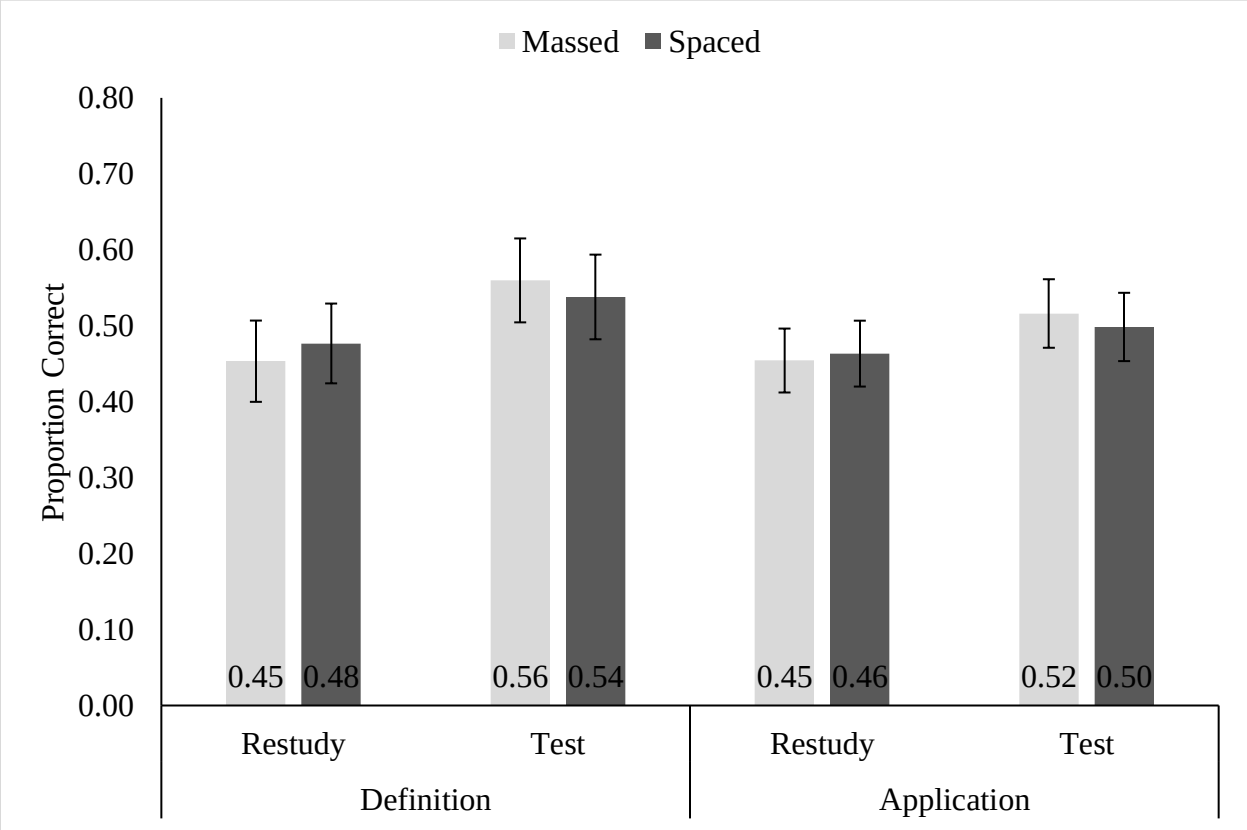

Figure 1 Performance of the Restudy and Test groups on massed and spaced items on definition and application questions in Experiment 1. Error bars indicate 95% confidence intervals.

The critical findings are displayed in Figures 1 and 2. Overall, participants performed slightly better on definition (M = 0.51, SD = 0.20) than on application questions (M = 0.48, SD = 0.15); however, this difference was only marginally significant, $F(1, 114) = 3.42$, $p = 0.07$, η_p^2 =

0.03. The Test group (M = 0.53, SD = 0.16) outperformed the Restudy group (M = 0.46, SD = 0.15), $F(1, 114) = 6.81$, $p = 0.01$, $\eta_p^2 = 0.06$—another replication of testing effects reported in the literature. Furthermore, this testing effect was observed for both definition and application questions (i.e., no significant interaction between review type and question type), $F(1, 114) = 1.80$, $p = 0.18$.

Interestingly, whether testing effects were observed depended on whether participants first saw definition or application questions on the criterial test (i.e., an interaction between review type and the counterbalancing variable of question order), $F(1, 114) = 3.71$, $p = 0.06$, $\eta_p^2 = 0.03$, though the effect was not large. Specifically, the Test group outperformed the Restudy group only when definition questions came first (Test group: M = 0.55, SD = 0.19, Restudy group: M = 0.43, SD = 0.13), $p = 0.002$, $d = 0.74$, but not when application questions came first (Test group: M = 0.50, SD = 0.13, Restudy group: M = 0.49, SD = 0.16), $p = 0.62$, $d = 0.07$. This pattern was similar for both definition and application questions; that is, the three-way interaction between review type, question type, and question order was not significant, $F(1, 114) = 0.04$, $p = 0.84$ (see Figure 2). To summarize, when participants answered definition questions first and then answered application questions, prior testing improved performance relative to restudy on both question types. When participants answered application questions first and then answered definition questions, testing did not improve performance relative to restudy for either question type.

Despite the robustness of spacing effects reported in the literature, criterial test performance was similar on spaced (M = 0.49, SD = 0.17) and massed concepts (M = 0.50, SD = 0.18), $F(1, 114) = 0.10$, $p = 0.75$. There was no interaction between review type and review timing, $F(1, 114) = 2.61$, $p = 0.11$, suggesting that the lack of a spacing effect was true for both

Test and Restudy groups. Similarly, neither definition nor application questions showed a spacing effect (i.e., no interaction between question type and review timing), $F(1, 114) = 0.06$, $p = 0.80$. The interaction between review timing (massed or spaced), review type (restudy or test), and question type (definition or application) was not significant either, $F(1, 114) = 0.19$, $p = 0.66$ (see Figure 1).

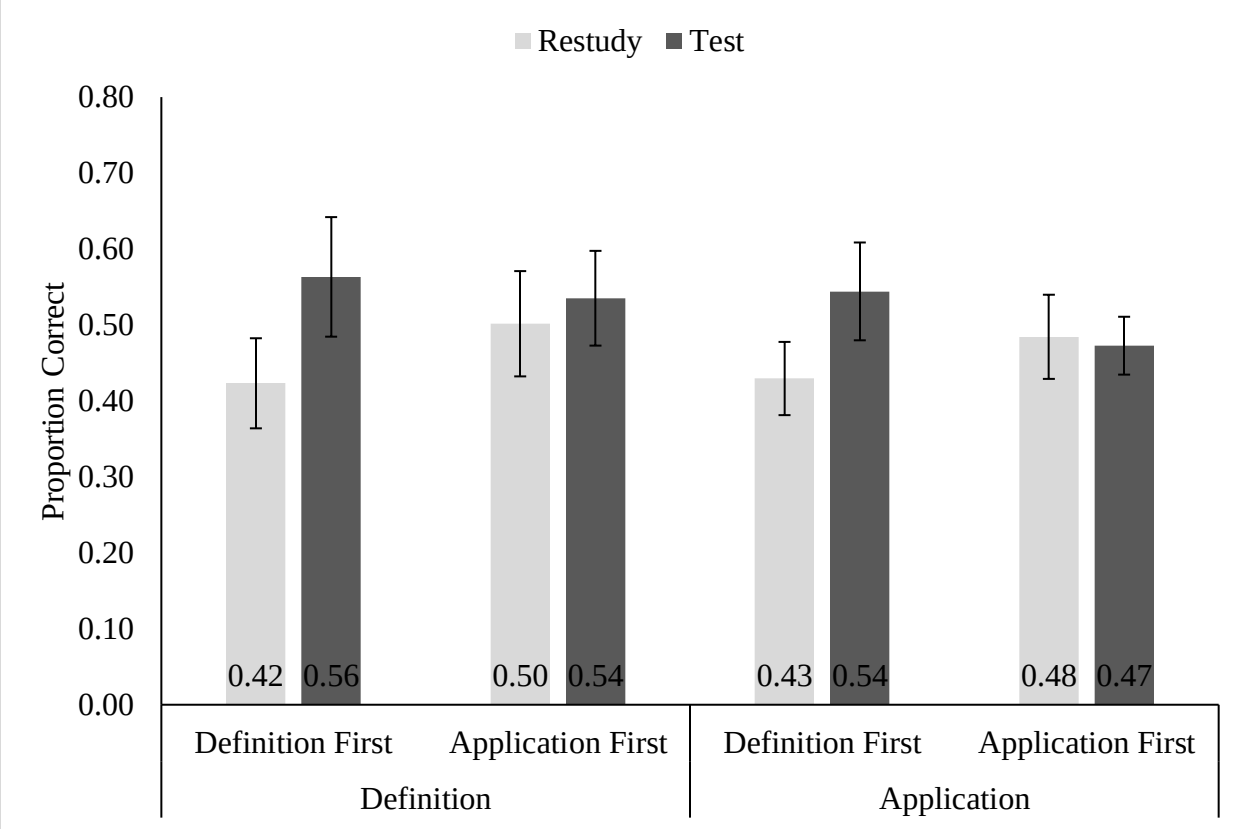

Figure 2 Performance of the Restudy and Test groups on definition and application questions based on which question type came first on the criterial test in Experiment 1. Definition First refers to the condition in which participants first answered definition and then application questions. Application First refers to the condition in which participants first answered application and then definition questions. Error bars indicate 95% confidence intervals.

Aside from the critical comparisons discussed above, two significant interactions emerged. There was a significant three-way interaction between review timing, which concept set was spaced, and question order, $F(1, 114) = 4.40$, $p = 0.04$, $\eta_p^2 = 0.04$. When definition questions came first on the criterial test and when concepts in Set 1 were spaced, there was a

marginal massing effect (p = 0.08); however, this pattern disappeared when concepts in Set 2 were spaced (p = 0.16). When application questions came first on the criterial test, by contrast, there were no differences between spaced and massed concepts based on which set was spaced (ps > 0.50). Yet, this interaction was qualified by a four-way interaction between review timing, review type, which concept set was spaced, and question order, $F(1, 114) = 4.98$, $p = 0.03$, $\eta_p^2 = 0.04$. This effect seems to be driven by the fact that the Test group showed a significant massing effect when Set 1 was spaced and when definition questions came first on the criterial test (p = 0.003). None of the other comparisons revealed a spacing or massing effect (all ps > 0.2). Both the three-way and four-way interactions involve two counterbalancing variables and are difficult to interpret; therefore, they are not discussed further.

2.2.2 Multi-Level Logistic Regression Results

Multi-level logistic regression was used to analyze criterial test responses (level 1) nested within participants (level 2), letting intercepts vary. Logistic regressions were conducted using generalized linear mixed-effects models from the *lme4* package (Bates et al., 2019) in RStudio (R Development Core Team, 2013). To analyze binary criterial test accuracy (0 for incorrect responses, 1 for correct responses), partially correct answers that were originally awarded 0.5 points were coded as 1 for lenient scoring. Lenient rather than strict scores were used to avoid floor effects, as there were very few fully correct responses in some questions, and a close look at partially correct responses revealed that most were qualitatively closer to correct responses than to incorrect responses.

To choose the best fitting model, I first created a model including all predictors—review type, review timing, question type, question order on the criterial test, and which concept set was spaced—and all possible two-way interactions between these predictors. I created a second

model which added all possible three-way interactions, and a third model that added all possible four-way interactions. A comparison of these nested models using a likelihood ratio test indicated that the three-way interaction model was better than the two-way interaction model, $x^2(10) = 23.93$, $p = 0.008$, but the three-way interaction and the four-way interaction models did not differ, $x^2(5) = 5.96$, $p = 0.31$. Thus, the more parsimonious model was selected. Any non-significant three-way and two-way interactions (all ps > 0.2) that were not relevant to the research questions were then removed to avoid over-fitting. This model, hereafter called Model 1, and the three-way interaction model were not different, $x^2(7) = 4.43$, $p = 0.73$; therefore, the more parsimonious model was selected. Model 1 is described in Table 2.

The results are presented in terms of odds ratios (*OR*), which refer to the change in the odds of correctly answering a criterial test question from the condition coded as 0 relative to the odds in the condition coded as 1. For example, if the Restudy group is coded as 0 and the Test group is coded as 1, an odds ratio greater than 1 would indicate a higher probability of correct recall in the Test than the Restudy group. By contrast, an odds ratio less than 1 would indicate a higher probability of correct recall in the Restudy than the Test group.

Table 2 Model 1 predicting criterial test accuracy for Experiment 1.

Predictors	Odds Ratio	CI	*p*
(Intercept)	0.81	0.55 – 1.21	0.3
Review Type	**2.18**	**1.39 – 3.41**	**0.001**
Review Timing	**1.62**	**1.14 – 2.29**	**0.01**
Question Type	1.15	0.82 – 1.61	0.43
Question Order	1.59	0.96 – 2.62	0.07
Spaced Concept Set	1.43	0.92 – 2.25	0.12
Review Type * Review Timing	0.78	0.55 – 1.10	0.15

Review Type * Question Type	0.87	0.57 – 1.33	0.52
Review Type * Question Order	0.62	0.35 – 1.09	0.1
Review Timing * Question Type	**0.58**	**0.39 – 0.88**	**0.01**
Review Timing * Question Order	1.34	0.94 – 1.90	0.11
Review Timing * Spaced Concept Set	**0.51**	**0.34 – 0.79**	**0.002**
Question Type* Question Order	0.83	0.59 – 1.17	0.2
Question Type * Spaced Concept Set	**0.71**	**0.50 – 0.99**	**0.05**
Question Order * **Spaced Concept Set**	0.95	0.54 – 1.68	0.87
Review Timing * Question Order * Spaced Concept Set	**0.54**	**0.33 – 0.88**	**0.01**
Review Type * Review Timing * Question Type	1.08	0.67 – 1.75	0.75
Review Type * Question Type * Question Order	0.92	0.57 – 1.49	0.73
Review Timing * Question Type * Spaced Concept Set	**2.43**	**1.50 – 3.94**	**<0.001**

Note. Model 1 includes all effects of interest regardless of whether they significantly predict criterial test accuracy. This model predicts criterial test accuracy (incorrect coded as 0, correct coded as 1) by review type (restudy coded as 0, test coded as 1), review timing (massed as 0, spaced as 1), question type (definition as 0, application as 1), question order on the criterial test (definition first as 0, application first as 1), which concept set was spaced (set 1 spaced as 0, set 2 spaced as 1), and interactions among these variables.

Model 1 includes all effects of interest; however, due to non-significant higher-order interactions, lower-order effects are difficult to interpret in this model. To obtain a more parsimonious and easily interpretable model, any non-significant three-way and two-way interactions (ps > 0.2) were further removed from Model 1, despite their relevance to the research questions. As such, a particular effect's absence from the model indicates that it was not significant. This model, hereafter called Model 2, was not different from the three-way interaction model ($x^2(10) = 6.01$, $p = 0.81$) or Model 1 ($x^2(3) = 1.59$, $p = 0.66$), and it is described in Table 3.

Table 3 Model 2 predicting criterial test accuracy for Experiment 1.

Predictors	Odds Ratio	CI	*p*
(Intercept)	0.84	0.57 – 1.23	0.36
Review Type	**2.03**	**1.37 – 3.01**	**<0.001**
Review Timing	**1.59**	**1.14 – 2.21**	**0.01**
Question Type	1.08	0.82 – 1.43	0.57
Question Order	1.62	1.00 – 2.64	0.05
Spaced Concept Set	1.45	0.92 – 2.26	0.11
Review Type * Review Timing	0.81	0.64 – 1.03	0.09
Review Type * Question Order	**0.59**	**0.35 – 0.99**	**0.05**
Review Timing * Question Type	**0.60**	**0.42 – 0.86**	**0.01**
Review Timing * Question Order	1.34	0.94 – 1.90	0.11
Review Timing * Spaced Concept Set	**0.51**	**0.34 – 0.79**	**0.002**
Question Type * Question Order	0.80	0.63 – 1.01	0.06
Question Type * Spaced Concept Set	**0.69**	**0.49 – 0.97**	**0.04**
Question Order * Spaced Concept Set	0.95	0.54 – 1.68	0.87
Review Timing * Question Order * Spaced Concept Set	**0.54**	**0.33 – 0.88**	**0.01**
Review Timing * Question Type * Spaced Concept Set	**2.45**	**1.51 – 3.96**	**<0.001**

Note. Model 2 only includes effects that significantly predict criterial test accuracy or effects that approach it (all *p*s < 0.2). This model predicts criterial test accuracy (incorrect coded as 0, correct coded as 1) by review type (restudy as 0, test as 1), review timing (massed as 0, spaced as 1), question type (definition as 0, application as 1), question order on the criterial test (definition first as 0, application first as 1), which concept set was spaced (set 1 spaced as 0, set 2 spaced as 1), and interactions among these variables.

Model 1 demonstrates that the critical two-way interaction between review type and question type and the two critical three-way interactions of theoretical relevance—one between review type, review timing, and question type; another between review type, question type, and question order—did not significantly predict criterial test accuracy, supporting the ANOVA

findings. In other words, neither the Test group nor the Restudy group demonstrated spacing effects, and these patterns were similar across definition and application questions. Furthermore, testing effects when definition or application questions came first did not vary across definition and application questions. Model 2 does not include these three interactions and is therefore the better model by virtue of parsimony.

The regression and ANOVA findings were quite similar. Thus, to avoid redundancy, only substantial discrepancies between the two analysis methods are discussed. According to the regression, there was a significant three-way interaction between review timing, question type, and which concept set was spaced ($OR = 2.45$, $p < 0.001$). This interaction occurred because, for application questions, there were no differences between massed and spaced concepts regardless of the concept set that was spaced. However, for definition questions, spacing led to better performance when Set 1 was spaced and massing led to better performance when Set 2 was spaced. This finding suggests that only concepts in Set 1 benefitted from spacing on the definition questions.

2.3 Discussion

Experiment 1 demonstrated conditions under which testing improves memory and novel application of concepts. Short-answer quizzes improved performance on the criterial test for both definition and application questions, but only when definition questions came first and application questions came second. That is, testing did not enhance performance on either question type when application questions preceded definition questions.

Arguably, examining the Test and Restudy groups' performance on definition questions when those came first and these groups' performance on application questions when those came first is the uncontaminated measure of test-enhanced retention and test-enhanced application,

respectively. With this consideration, short-answer definition quizzes improved memory only for concept definitions (14% difference between Test and Restudy groups), but not for novel application of concepts (1% difference between Test and Restudy groups). Though the retention benefits of initial quizzing replicate prior testing effects (see Rowland, 2014, for a meta-analysis), the finding that quizzes did not improve performance on application questions contrasts prior studies demonstrating test-enhanced transfer (see Pan & Rickard, 2018, for a meta-analysis).

Interestingly, however, the above pattern reversed when performance from the second half of the criterial test was examined. Specifically, the Test group outperformed the Restudy group on the application questions by about 11% when both groups answered definition questions prior—a testing effect on application questions (Butler, 2010). Thus, the evidence for test-enhanced transfer is mixed in Experiment 1, with quizzing improving application question performance over restudy only when application questions followed definition questions, but not when application questions were first on the criterial test.

In addition, the Test group performed only about 4% (and not significantly) better than the Restudy group on the definition questions when both groups first answered application questions. It is possible that the act of retrieval for the Restudy group—in the form of answering application questions on the first half of the criterial test—improved their performance on the following definition questions, which could have closed the gap between Test and Restudy groups on definition questions. In fact, Restudy participants who saw definition questions first performed at 42% on definition questions, whereas Restudy participants who saw definition questions second performed at 50% on these definition questions. However, this difference was not reliable. It is also possible that, due to chance, Restudy participants who were assigned to the

condition that first saw application questions performed better overall than Restudy participants who were assigned to the condition that first saw definition questions.

In Experiment 1, the lack of spacing effects, for both groups and on both question types, was surprising and contrary to prior research showing robust spacing benefits (Cepeda et al., 2016). Spacing effects were expected particularly on definition questions; however, massed and spaced conditions were almost identical. There are several possible explanations for why spacing effects were not observed. First, when participants reviewed the concepts massed, they did not see items back-to-back (quiz questions for the Test group, concept definitions for the Restudy group), but they saw repetitions nine items apart. That is, rather than particular items being massed, the whole quiz or restudy was massed. Other items introduced between the first and second presentation of an item could have eliminated spacing effects; however, prior studies on lag effects suggest that longer lags should lead to greater retention than shorter lags, and the difference between nine items (massed) and two days (spaced) is quite large. Given that overall performance on massed and spaced concepts was almost identical, this explanation alone seems unlikely.

Additionally, the lack of a spacing effect could be explained by the relationship between the spacing gap and the retention interval (two days for both), as the prior literature suggests a particular spacing gap may be optimal for a given retention interval (Cepeda et al., 2008). It is possible that the retention interval used in Experiment 1 may not have been long enough, and a longer retention interval, for instance, of one week could have revealed significant spacing effects on definition questions (see Pyc et al., 2014, for similar considerations)

Furthermore, the online nature of data collection might explain the lack of spacing effects. Specifically, participants who completed the experiment online (more than half of the

participants) were sent a link for a follow-up session on the day they were supposed to complete it. Although participants were instructed to start all three sessions at the same time, this was not always the case; participants' start times varied across sessions. Nonetheless, even when only considering participants who completed the experiment in the laboratory, where sessions were 48 hours apart, performance on massed definition questions (M = 0.47, SD = 0.20) and spaced definition questions (M = 0.46, SD = 0.21) were still identical.

Table 4 Performance of the Test group on the quizzes and criterial test's definition questions for massed and spaced concepts in Experiment 1.

	Quiz 1	Quiz 2	Criterial Test
Massed	.33 *(.18)*	.52 *(.21)*	.56 *(.22)*
Spaced	.38 *(.16)*	.40 *(.17)*	.53 *(.22)*

Note. Standard deviations are reported in parentheses. Only performance from the definition questions is displayed for the criterial test. Quizzes and the criterial test contain short-answer questions.

Finally, one issue to consider is initial quiz differences between massed and spaced concepts for the Test group. Because the Test group had greater Quiz 2 accuracy for massed than spaced concepts, it is possible that this difference masked any spacing effects that could have been observed on the criterial test. On the other hand, correct-answer feedback provided on the second quiz should have equated performance to some extent. One way to examine this issue is presented in Table 4. Performance of the Test group on the two quizzes and the criterial test (for definition questions), split by massed and spaced items, reveals the gain between the second quiz and the criterial test. The Test group's performance on massed concepts increased by 4% from the second quiz to the criterial test, whereas this increase was 13% for spaced concepts. A 2 (quiz 2, criterial test) x 2 (massed, spaced) within-subjects ANOVA revealed that this interaction was significant ($F(1, 60) = 24.83$, $p < 0.001$, $\eta_p^2 = 0.29$); performance increased only marginally from the second quiz to the criterial test for massed concepts (4%, $p = 0.09$, $d = 0.22$), whereas

this increase was significant for spaced concepts (13%, $p < 0.001$, $d = 0.98$). Though criterial test performance did not differ on massed and spaced concepts, the differences in performance change from the last quiz to the criterial test suggest that spacing and massing operate differently to yield similar subsequent test performances.

Taken together, Experiment 1's findings provide some boundary conditions to when testing effects are observed on verbatim retention and novel application after students have learned from a textbook. However, because spacing effects were not observed in Experiment 1, the primary question of interest—whether there are combined benefits of testing and spacing on novel application questions—could not be evaluated. In Experiment 2, multiple-choice initial quizzes were employed to potentially diminish initial quiz differences between massed and spaced items, and to test the reliability of overall findings from Experiment 1.

Chapter 3: Experiment 2

The goal of Experiment 2 was to examine the reliability of Experiment 1's findings with a different initial quiz format. Instead of the short-answer quizzes used in Experiment 1, multiple-choice quizzes were used in Experiment 2 for several reasons. First, using multiple-choice quizzes was predicted to increase quiz accuracy of the Test group. Quiz accuracy, which is a proxy of initial learning, is an important factor in the presence and magnitude of testing benefits (Karpicke, 2017; Rowland, 2014). In Experiment 1, even with the correct-answer feedback presented on the first quiz, the Test group had about 46% accuracy on the second quiz, which consisted of questions identical to the first quiz. Furthermore, even though the second quiz for concepts in the massed condition occurred only several minutes after the first quiz, the Test group performed at 52%. Though feedback was provided after the second quiz as well, performance on the second quiz suggested that the Test group could have gained more knowledge about the concepts. By using multiple-choice quizzes in Experiment 2, quiz accuracy was expected to be greater relative to that in Experiment 1 with short-answer quizzes. This, in turn, was predicted to result in more robust testing effects on the criterial test.

Another reason multiple-choice quizzes were used in Experiment 2 was to diminish the performance difference between massed and spaced items on the quizzes. In Experiment 1, on the second quiz, participants performed better on the concepts in the massed condition compared to concepts in the spaced condition. As mentioned above, this difference could have masked spacing effects on the criterial test for the Test group. With multiple-choice quizzes, it was predicted that participants would more easily remember the correct answer from the feedback when a particular question was repeated, as opposed to generating the correct answer on a short-answer question. As such, the use of multiple-choice questions in Experiment 2 also served to

diminish and possibly eliminate the quiz difference between massed and spaced concepts. If so, it was predicted for the Test group to show significant spacing effects on the criterial test, particularly on the definition questions.

Finally, using multiple-choice quizzes allowed testing the generalizability of Experiment 1's findings in more common classroom settings. It is often the case that students have multiple-choice practice questions available when they study for upcoming exams, even though the exams may consist of short-answer questions. In Experiment 1, quizzes as well as criterial tests consisted of short-answer questions. In Experiment 2, however, participants in the Test group took multiple-choice quizzes instead of short-answer quizzes. The criterial test still consisted of short-answer questions.

3.1 Method

3.1.1 Participants

The same power analysis from Experiment 1 was used to set the sample size. 112 Washington University undergraduates (M_{age} = 19.64, SD_{age} = 1.22) participated in all three sessions of the experiment online for course credit (2.5 credits) or payment ($20). To incentivize participants to complete follow-up sessions, participants received a $5 bonus payment if they completed all sessions. Seven additional participants started the experiment; three did not return for Session 2 and four did not return for Session 3.

3.1.2 Materials, Design, and Procedure

The materials, design, and procedure of Experiment 2 were identical to Experiment 1 except for the following changes: Participants in the Test group in Experiment 2 took multiple-choice quizzes, rather short-answer quizzes, and four-option multiple-choice definition questions were

created to use in these quizzes. As an example, the short-answer definition question for *repeated-measures design* was changed from "What is a repeated-measures design?" to the following:

What is a repeated-measures design?

(a) Each participant is measured under all conditions of an independent variable

(b) A different group of participants is randomly selected for each condition of an independent variable

(c) Each participant is matched with a participant in every other condition based on an extraneous variable

(d) An experiment is replicated using the same variables and measures

In addition, the experiment was conducted fully online. However, unlike Experiment 1, an experimenter was present during Session 1 of the experiment, as participants had to log onto a Zoom meeting and keep their video on for the duration of the experiment. This allowed more control of participants' behavior and environment for the longest session of the experiment. Sessions 2 and 3 were conducted similarly to Experiment 1, where participants received an email with an experiment link on the day of the follow-up sessions. The methodology of Experiment 2 was otherwise identical to that of Experiment 1.

3.1.3 Scoring

Multiple-choice quizzes were automatically scored by the computer. Short-answer responses on the criterial tests were hand-scored, where correct answers were given 1 point, partially correct answers were given 0.5 points, and incorrect answers were given 0 points based on the same rubric as Experiment 1. As raters showed good inter-rater reliability in Experiment 1, the same rater scored all responses in Experiment 2.

3.2 Results

112 participants completed all three sessions of the experiment (n = 55 for the Restudy group, n = 57 for the Test group). Three of these participants (all in the Test group) indicated taking notes while reading the chapter; however, whether these participants were included or not did not change the findings. Thus, these participants are included in the analyses. Fifteen participants who completed all sessions (eight in the Restudy group and seven in the Test group) indicated that they previously took or that they were currently taking statistics or research methods courses. The analyses include these participants as well, as excluding them did not change the findings.

As in Experiment 1, ACT scores (range: 25-36, M = 33.26, SD = 2.28) were used to determine if those who completed the experiment differed from those who did not return for either follow-up session. Three participants did not return for Session 2 and four did not return for Session 3. As very few participants reported SAT scores alone, these were converted to ACT scores. Of the participants who reported ACT scores, those who did not return for Sessions 2 or 3 (n = 5, M = 32.20, SD = 2.17) were similar to those who completed all sessions (n = 106, M = 33.26, SD = 2.28), and this was the case when split between Restudy and Test groups as well.

Similar to Experiment 1, I first report quiz and criterial test results using the ANOVA framework. All omnibus tests of statistical significance use an alpha level of 0.05 and all pairwise comparisons are reported with a Bonferroni correction. Effect sizes are reported using partial eta-squared for main effects and interactions, and Cohen's d for pairwise comparisons. Similarly, I also report results using multi-level logistic regression as a complementary analysis, to better account for within-person correlations and examine trial-level accuracy. Again, I focus

on the ANOVA results for ease of interpretation. The regression models are presented in tables, but only patterns that differed from the ANOVAs are described to avoid redundancy.

3.2.1 ANOVA Results

Quiz Performance. During Sessions 1 and 2, the Test group took multiple-choice quizzes with feedback. Participants completed one quiz twice in a massed fashion in Session 1 and completed another quiz twice in a spaced fashion across Sessions 1 and 2. A 2 (quiz number: quiz 1 or quiz 2) x 2 (review timing: massed or spaced) x 2 (which concept set was spaced: set 1 or set 2) mixed factorial ANOVA was conducted to examine quiz performance. The order questions appeared on the quizzes was included at first, but it was dropped as it did not have a main effect and did not interact with any of the other variables (all *p*s > 0.1).

Quiz performance is displayed in Table 5. Unsurprisingly, given the provision of correct-answer feedback after each quiz question, Quiz 2 performance (M = 0.81, SD = 0.14) was better than Quiz 1 performance (M = 0.67, SD = 0.13), $F(1, 55) = 71.57$, $p < 0.001$, $\eta_p^2 = 0.57$. Participants also performed better overall on massed (M = 0.77, SD = 0.12) than spaced concepts (M = 0.71, SD = 0.16), $F(1, 55) = 9.34$, $p = 0.003$, $\eta_p^2 = 0.15$. Critically, there was a significant interaction between review timing and quiz number, $F(1, 55) = 29.10$, $p < 0.001$, $\eta_p^2 = 0.35$. Although performance on massed (M = 0.66, SD = 0.16) and spaced concepts (M = 0.68, SD = 0.18) did not differ on Quiz 1 ($p = 0.41$, $d = 0.15$), massed concepts (M = 0.88, SD = 0.14) were recalled better than spaced concepts (M = 0.73, SD = 0.19) on Quiz 2 ($p < 0.001$, $d = 0.81$). Put differently, performance from Quiz 1 to Quiz 2 improved more for the massed concepts (by 22%) than for the spaced concepts (by 5%), likely due to being able to hold the feedback in mind when answering the same quiz question in close succession. These findings are similar to those from Experiment 1. However, the finding that performance on massed concepts was better than

spaced concepts on the last quiz contrasts with the prediction that this difference would have decreased or disappeared when switching from the short-answer quizzes in Experiment 1 to the multiple-choice quizzes in this experiment.

Table 5 Performance of the Test group on Quiz 1 and Quiz 2 on massed and spaced concepts in Experiment 2.

	Quiz 1	Quiz 2
Massed	.66 *(.16)*	.88 *(.14)*
Spaced	.68 *(.18)*	.73 *(.19)*

Note. Standard deviations are reported in parentheses.

The counterbalancing variable of which concept set was spaced did not have a main effect and it did not interact with quiz number (both *p*s > 0.1); however, this variable had a significant interaction with review timing, $F(1, 55) = 14.69$, $p < 0.001$, $\eta_p^2 = 0.21$. This interaction occurred because when concepts in Set 1 were massed (and concepts in Set 2 were spaced), participants did better on massed ($M = 0.79$, $SD = 0.13$) than spaced ($M = 0.65$, $SD = 0.16$) concepts ($p < 0.001$, $d = 0.96$). However, when concepts in Set 2 were massed (and concepts in Set 1 were spaced), participants performed similarly on massed ($M = 0.74$, $SD = 0.11$) and spaced ($M = 0.76$, $SD = 0.15$) concepts ($p = 0.58$, $d = 0.15$). In other words, quiz accuracy for Set 1 concepts did not change much whether they were scheduled as massed ($M = 0.79$) or spaced ($M = 0.76$). However, quiz accuracy for Set 2 concepts did differ based on whether they were massed ($M = 0.74$) or spaced ($M = 0.65$). Five out of ten quiz questions for Set 2 concepts asked participants for the concept name given the definition, whereas only one out of ten questions for Set 1 concepts did so (all other questions asked for the definition given the concept name). It is possible that these questions were easier overall, and participants improved their performance even more based on the correct-answer feedback at the time of Quiz 2.

However, the three-way interaction between which concept set was spaced, review timing, and quiz number was not significant, $F(1, 55) = 0.85$, $p = 0.36$.

Criterial Test Performance. A 2 (review type: restudy or test) x 2 (review timing: massed or spaced) x 2 (question type: definition or application) x 2 (which concept set was spaced: set 1 or set 2) x 2 (question order: definition first or application first) mixed factorial ANOVA was conducted to examine criterial test performance. Neither of the counterbalancing variables had a main effect on performance (both *p*s > 0.1). However, because these variables interacted with the main variables, they were not taken out of the ANOVA.

Criterial test results are displayed in Figures 3 and 4. Similar to Experiment 1, participants performed slightly better on definition ($M = 0.53$, $SD = 0.16$) than application questions ($M = 0.50$, $SD = 0.14$), $F(1, 104) = 4.52$, $p = 0.04$, $\eta_p^2 = 0.04$. The 3% difference between definition and application questions is the same as in Experiment 1, though the difference was only marginally significant in Experiment 1.

Similar to Experiment 1 and in line with prior testing effects, the Test group ($M = 0.55$, $SD = 0.12$) outperformed the Restudy group ($M = 0.49$, $SD = 0.15$), $F(1, 104) = 4.91$, $p = 0.03$, $\eta_p^2 = 0.05$. Unlike Experiment 1, however, there was a significant interaction between review type and question type, $F(1, 104) = 9.73$, $p = 0.002$, $\eta_p^2 = 0.09$. Specifically, the Test group ($M = 0.58$, $SD = 0.14$) outperformed the Restudy group ($M = 0.48$, $SD = 0.17$) on definition questions ($p = 0.002$, $d = 0.64$); however, the Test ($M = 0.52$, $SD = 0.13$) and Restudy ($M = 0.49$, $SD = 0.16$) groups performed similarly on application questions ($p = 0.43$, $d = 0.21$). Put differently, multiple-choice quizzes did not improve students' ability to apply their concept knowledge above restudying concept definitions.

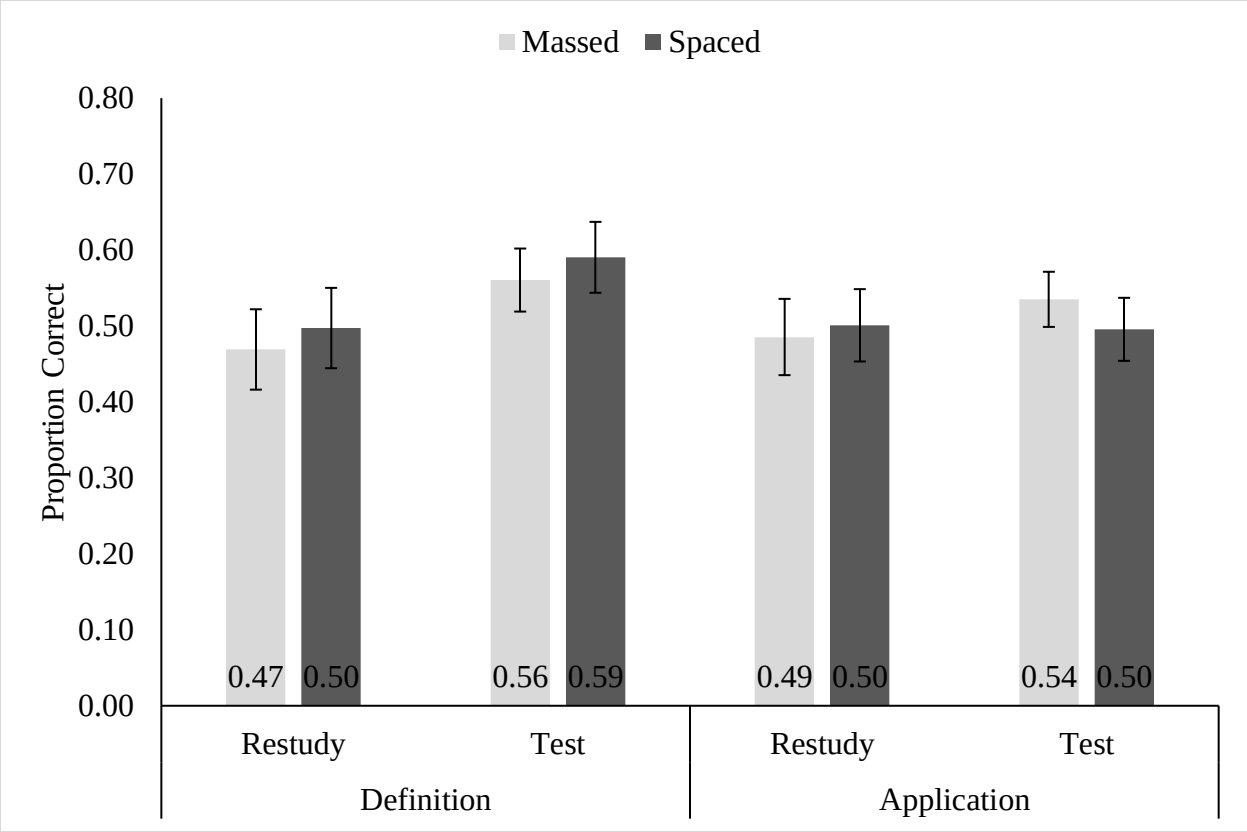

Figure 3 Performance of the Restudy and Test groups on massed and spaced items on definition and application questions in Experiment 2. Error bars indicate 95% confidence intervals.

Unlike Experiment 1, the presence of testing effects did not depend on whether definition or application questions came first on the criterial test (i.e., no significant interaction between question order and review type), $F(1, 104) = 0.99$, $p = 0.32$. This pattern also did not change based on whether the questions were definition or application questions (i.e., no significant interaction between question order, question type, and review type), $F(1, 104) = 0.05$, $p = 0.82$ (see Figure 4).

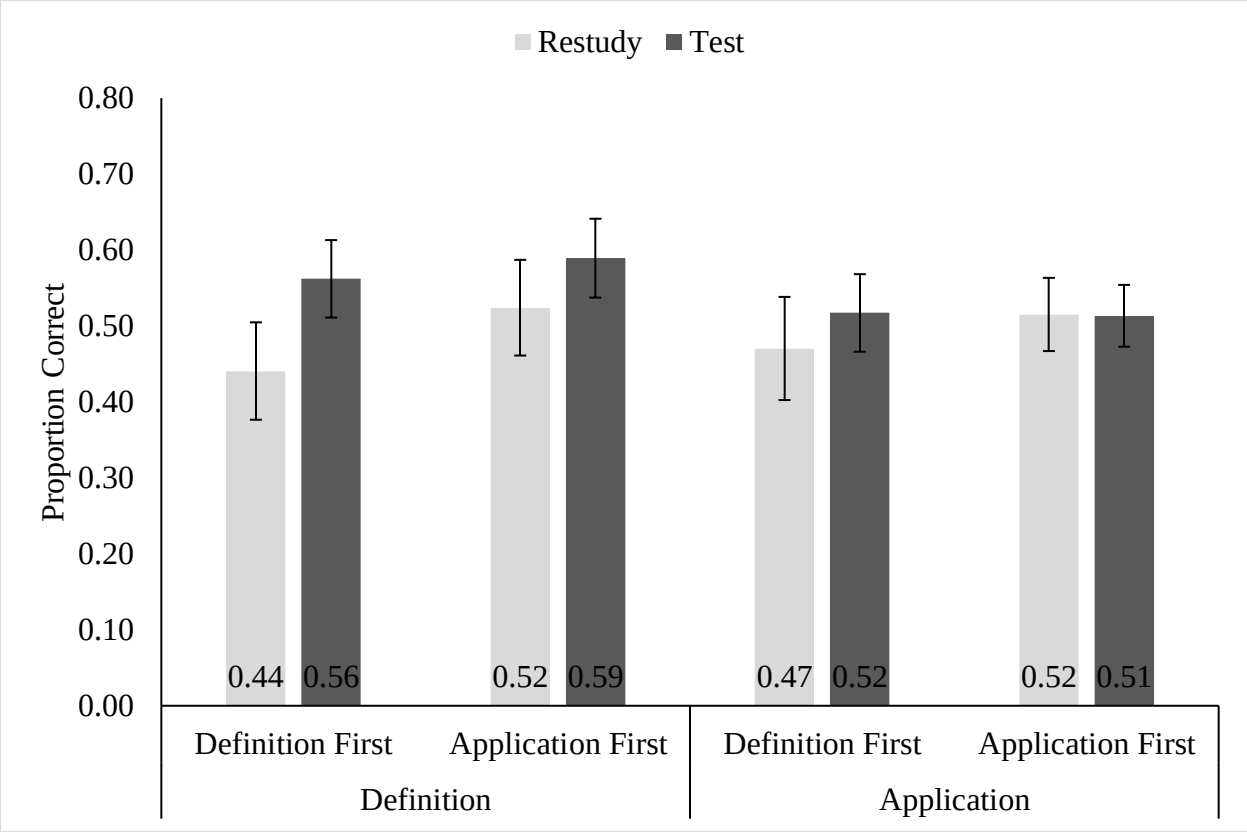

Figure 4 Performance of the Restudy and Test groups on definition and application questions based on which question type came first on the criterial test in Experiment 2. Definition First refers to the condition in which participants first answered definition and then application questions. Application First refers to the condition in which participants first answered application and then definition questions. Error bars indicate 95% confidence intervals.

Regarding spacing effects on learning, Experiment 2 findings replicated Experiment 1; that is, criterial test performance on concepts reviewed spaced (M = 0.52, SD = 0.16) or massed (M = 0.51, SD = 0.16) did not differ, $F(1, 104) = 0.40$, $p = 0.53$. Again, similar to Experiment 1, neither the Test nor the Restudy group benefitted from spaced review of concepts (i.e., no interaction between review type and review timing), $F(1, 104) = 0.99$, $p = 0.32$. Also as in Experiment 1, there were no interactions between review timing, review type, and question type, $F(1, 104) = 2.04$, $p = 0.16$ (see Figure 3).

However, unlike Experiment 1, there was a marginally significant interaction between review timing and question type, $F(1, 104) = 3.71$, $p = 0.06$, $\eta_p^2 = 0.03$. This interaction occurred

because for massed concepts, there were no performance differences between definition (M = 0.52, SD = 0.19) and application questions (M = 0.51, SD = 0.17), p = 0.77, d = 0.03. However, for spaced concepts, performance was better on definition (M = 0.54, SD = 0.19) than on application questions (M = 0.50, SD = 0.17), p = 0.005, d = 0.25. Furthermore, this pattern differed based on which concept set was spaced or massed (i.e., a significant three-way interaction occurred between review timing, question type, and the concept set that was spaced), $F(1, 104) = 4.65$, $p = 0.03$, $\eta_p^2 = 0.04$. When Set 1 was spaced and Set 2 was massed, there was a significant spacing effect on the definition questions (M = 0.56, SD = 0.18 for spaced, and M = 0.50, SD = 0.21 for massed, p = 0.02, d = 0.31); however, there were no spacing effects on any of the application questions when either set was spaced and there was no spacing effect on definition questions when Set 2 was spaced.

The above patterns were qualified by a four-way interaction between review timing, question type, spaced concept set, and review type, $F(1, 104) = 8.02$, $p = 0.01$, $\eta_p^2 = 0.07$. A close look at spacing effects for each group (Test or Restudy), for each question type (definition or application), and under each counterbalancing condition (Set 1 spaced or Set 2 spaced) revealed that the pattern in the above paragraph was only observed for the Restudy group. That is, a significant spacing effect was observed for the Restudy group on definition questions only when concepts in Set 1 were spaced (M = 0.54, SD = 0.19 for spaced, and M = 0.44, SD = 0.23 for massed, p = 0.01, d = 0.47). No other conditions showed spacing effects on the criterial test (all ps > 0.1).

In addition to the effects reported above, there were three marginally significant four-way interactions, and one marginally significant three-way interaction. Because these interactions included at least one counterbalancing variable, did not pertain to the critical research questions,

were only marginally significant, and were difficult interpret due to being three-way or four-way interactions, they are not discussed. The description of these interactions can be found in Appendix C.

3.2.2 Multi-Level Logistic Regression Results

Multi-level logistic regression was used to analyze criterial test responses (level 1) nested within participants (level 2), letting intercepts vary. Logistic regressions were conducted using generalized linear mixed-effects models from the *lme4* package (Bates et al., 2019) in RStudio (R Development Core Team, 2013). To analyze binary criterial test accuracy (0 points for incorrect responses, 1 for correct responses), partially correct answers were again coded as 1 for lenient scoring, for reasons described in Experiment 1.

To select the best fitting model, the same process from Experiment 1 was used. Three nested models based on the same fixed effects (review type, review timing, question type, question order on the criterial test, and which concept set is spaced)—one adding all possible two-way interactions, another adding all possible three-way interactions, and another adding all possible four-way interactions—were compared using a likelihood ratio test. The three-way interaction model was better than the two-way interaction model, $x^2(10) = 29.16$, $p = 0.001$, but not better than the four-way interaction model, $x^2(5) = 9.66$, $p = 0.09$. Thus, the parsimonious three-way interaction model was selected. Any non-significant three-way and two-way interactions (all ps > 0.2) not relevant to the research questions were then removed to avoid over-fitting, as in Experiment 1. A comparison between this model, hereafter referred to as Model 3, and the full three-way interaction model showed no differences between the two, $x^2(10) = 5.95$, $p = 0.82$. Model 3 is described in Table 6.

Table 6 Model 3 predicting criterial test accuracy for Experiment 2.

Predictors	Odds Ratios	CI	*p*
(Intercept)	0.76	0.55 – 1.06	0.1
Review Type	**1.82**	**1.21 – 2.75**	**0.004**
Review Timing	**2.49**	**1.82 – 3.39**	**<0.001**
Question Type	1.35	0.95 – 1.91	0.09
Question Order	**1.61**	**1.11 – 2.34**	**0.01**
Spaced Concept Set	**2.05**	**1.49 – 2.82**	**<0.001**
Review Type * Review Timing	1.00	0.70 – 1.44	0.99
Review Type * Question Type	0.77	0.50 – 1.19	0.24
Review Type * Question Order	0.71	0.42 – 1.21	0.21
Review Timing * Question Type	**0.49**	**0.32 – 0.75**	**0.001**
Review Timing * Spaced Concept Set	**0.22**	**0.15 – 0.31**	**<0.001**
Question Type * Question Order	0.79	0.55 – 1.12	0.18
Question Type * Spaced Concept Set	**0.63**	**0.44 – 0.90**	**0.01**
Review Type * Review Timing * Question Type	0.77	0.47 – 1.28	0.31
Review Type * Question Type * Question Order	1.12	0.68 – 1.85	0.65
Review Timing * Question Type * Spaced Concept Set	**3.54**	**2.14 – 5.85**	**<0.001**

Note. Model 3 includes all effects of interest regardless of whether they significantly predict criterial test accuracy. This model predicts binary criterial test accuracy (incorrect as 0, correct as 1) by review type (restudy as 0, test as 1), review timing (massed as 0, spaced as 1), question type (definition as 0, application as 1), counterbalancing of question order on the criterial test (definition first as 0, application first as 1), counterbalancing of which concept set was spaced (set 1 spaced as 0, set 2 spaced as 1), and interactions among these variables.

Similar to the model selection process in Experiment 1, an even more parsimonious model, hereafter referred to as Model 4, was created by taking out non-significant two-way and three-way interactions (all *p*s > 0.2) that were initially included for their relevance to the research

questions. This pruning allows for easier interpretation of the model's lower-order effects, and Model 4 was not different from the full three-way interaction model ($x^2(13) = 7.87$, $p = 0.85$) or Model 3 ($x^2(3) = 1.92$, $p = 0.59$). Model 4 is described in Table 7. The absence of a particular effect from Model 4 indicates that the effect was not significant.

Table 7 Model 4 predicting criterial test accuracy for Experiment 2.

Predictors	Odds Ratios	CI	*p*
(Intercept)	0.89	0.65 – 1.21	0.46
Review Type	**1.54**	**1.19 – 2.01**	**0.001**
Review Timing	**2.48**	**1.92 – 3.21**	**<0.001**
Question Type	**1.40**	**1.04 – 1.89**	**0.03**
Question Order	1.17	0.83 – 1.65	0.38
Spaced Concept Set	**1.76**	**1.19 – 2.60**	**0.01**
Review Type * Question Type	**0.72**	**0.56 – 0.92**	**0.01**
Review Timing * Question Type	**0.43**	**0.30 – 0.61**	**<0.001**
Review Timing * Spaced Concept Set	**0.22**	**0.15 – 0.31**	**<0.001**
Question Type * Question Order	0.83	0.65 – 1.07	0.15
Question Type * Spaced Concept Set	**0.63**	**0.44 – 0.90**	**0.01**
Question Order * Spaced Concept Set	1.36	0.86 – 2.15	0.19
Review Timing * Question Type * Spaced Concept Set	**3.53**	**2.14 – 5.84**	**<0.001**

Note. Model 4 only includes effects that significantly predict criterial test accuracy or effects that approach it (all *p*s < 0.2). This model predicts binary criterial test accuracy (incorrect as 0, correct as 1) by review type (restudy as 0, test as 1), review timing (massed as 0, spaced as 1), question type (definition as 0, application as 1), counterbalancing of question order on the criterial test (definition first as 0, application first as 1), counterbalancing of which concept set was spaced (set 1 spaced as 0, set 2 spaced as 1), and interactions among these variables.

Replicating the model from Experiment 1, Model 3 demonstrates that the critical two-way interaction between review type and question type and the two critical three-way interactions of theoretical relevance—one between review type, review timing, and question

type; another between review type, question type, and question order—did not reliably predict criterial test accuracy. Model 4 does not include these interactions (as well as other non-significant effects), and it is therefore the better model due to its parsimony.

Similar to Experiment 1, the regression findings largely aligned with the ANOVA results. Thus, to avoid redundancy, only the discrepancies between the two methods are discussed. Based on Model 4, there was a significant three-way interaction between review timing, question type, and concept set that was spaced ($OR = 3.53$, $p < 0.001$). These findings are the same with findings from Experiment 1's regression, and furthermore, this three-way interaction was also significant in the ANOVA. Similar to the ANOVA, the regression showed a significant spacing effect on the definition questions when concepts in Set 1 were spaced. Unlike the pairwise comparisons in the ANOVA, with lenient scoring for the regression, when concepts in Set 2 were spaced, there was a massing effect on the definition questions (no differences on application questions were observed).

Finally, the regression did not bear out a marginally significant interaction from the ANOVA (between question type, review type, concept set that was spaced), which was taken out of the model in the model selection process. It is possible that the lenient scoring used for the regression, or the greater variance captured by the regression model at the trial level resulted in the discrepancy.

3.3 Discussion

One key finding from Experiment 2 was that taking multiple-choice quizzes with feedback improved criterial test performance on definition questions, but not on novel application questions, relative to restudying concept definitions. The finding that multiple-choice quizzes improved retention of concepts over restudying replicates a vast body of research demonstrating

testing effects (Rowland, 2014). Furthermore, the fact that multiple-choice quizzes yield retention benefits over restudying on a short-answer criterial test suggests that the match in test format between quizzes and criterial tests does not matter (Pan & Rickard, 2018).

The finding that multiple-choice definition quizzes did not improve performance on application questions relative to restudying concept definitions is contrary to studies demonstrating test-enhanced transfer (Pan & Rickard, 2018). However, many of the studies that demonstrated testing effects on the kinds of application and inference questions used in this experiment employed free recall or short-answer quizzes (see Table 1 from Pan & Rickard, 2018). The few studies that did use multiple-choice quizzes and obtained transfer to application- and inference-based questions did not quiz participants on concept definitions or facts, but quizzed them on the kinds of questions that appeared on the criterial test (e.g., Agarwal, 2019; McDaniel et al., 2015). In fact, Agarwal (2019) showed that multiple-choice fact quizzes did not improve performance on higher-order criterial tests; rather, only higher-order quizzes improved transfer. If multiple-choice quizzes in Experiment 2 asked participants to apply their concept knowledge, rather than recall concept definitions, test-enhanced transfer might have been observed on the criterial test.

As in Experiment 1, spacing effects were not obtained for either the Restudy group or the Test group. However, spacing effects emerged under a particular condition in Experiment 2: When Set 1 concepts were spaced, the Restudy group performed better on spaced definition questions than they did on massed definition questions. However, the Restudy group performed similarly on spaced and massed definition questions when Set 2 concepts were spaced. In Set 1, questions mostly asked participants for a definition given the concept name, except for one question (out of ten), whereas in Set 2, half the questions asked for the concept name given the

definition. Perhaps questions that asked for the concept name were already easy for participants (i.e., Set 2 concepts) and did not require spaced review to boost performance on those questions. On the other hand, questions that required generating the full definition were presumably more difficult and thus might have benefitted from spaced review. Nonetheless, the fact that spacing effects could not be generalized to the full set of material, to the Test group, and to application questions points towards the elusiveness of spacing effects in this paradigm.

The potential explanations as to why spacing effects were not observed in Experiment 2 are identical to those outlined for Experiment 1. First, the fact that the massed condition had a slight spacing between item repetitions could have mitigated spacing effects. Second, the retention interval might not have been long enough to yield spacing effects. Third, initial quiz differences between massed and spaced items for the Test group might have masked spacing effects for this group. However, Experiment 2 was not designed to address these issues, and the lack of spacing effects in both experiments are addressed in the General Discussion.

Table 8 Performance of the Test group on the quizzes and criterial test's definition questions for massed and spaced concepts in Experiment 2.

	Quiz 1	Quiz 2	Criterial Test
Massed	.66 *(.16)*	.88 *(.14)*	.56 *(.16)*
Spaced	.68 *(.18)*	.73 *(.19)*	.59 *(.18)*

Note. Standard deviations are reported in parentheses. Only performance from the definition questions is displayed for the criterial test. Quizzes contained multiple-choice questions, and the criterial test contained short-answer questions.

Similar to Experiment 1, the Test group had better performance on massed than spaced concepts on the second quiz, in contrast to the prediction that this difference would diminish or disappear in Experiment 2 with multiple-choice quizzes. Though the multiple-choice quiz performance in Experiment 2 was greater overall than the short-answer quiz performance in Experiment 1, the difference between massed and spaced quiz items was similar to that observed

in Experiment 1. Though contrary to predictions, this difference could have influenced spacing effects on the criterial test. Because the quizzes were in multiple-choice format and the criterial test was in short-answer format, overall performance dropped from the former to the latter. Table 8 shows that the Test group's performance on massed definition questions dropped by 32%, whereas this group's performance on spaced definition questions dropped by 14%. Similar to Experiment 1, a 2 (quiz 2, criterial test) x 2 (massed, spaced) within-subjects ANOVA revealed that this interaction was significant ($F(1, 56) = 27.01$, $p < 0.001$, $\eta_p^2 = 0.33$); performance dropped significantly from the second quiz to the criterial test for massed concepts (32%, $p < 0.001$, $d = 1.69$) and spaced concepts (14%, $p < 0.001$, $d = 0.77$), but this drop was greater for the former. Though there were no differences between spaced and massed concepts on the criterial test, the change from the second quiz to the criterial test suggests that spaced tests could slow forgetting relative to massed tests.

Chapter 4: General Discussion

The impetus of this dissertation was to investigate whether the combined benefits of distributed practice and retrieval practice observed on verbatim retention in prior research would also extend to situations in which students need to use their knowledge for novel application and inference. To address this issue, the current study examined students' learning of key concepts from a textbook—both memory and novel application—by varying whether students reviewed concepts massed or spaced as well as whether they reviewed concepts via restudy or quizzes with feedback. In two experiments, college students read from a research methods textbook, and they reviewed key concepts from it in one of four ways: massed restudy, spaced restudy, massed quizzes, or spaced quizzes. Students who restudied simply read definitions of key concepts from the textbook, whereas quizzed students answered short-answer (Experiment 1) or multiple-choice (Experiment 2) definition questions and received correct-answer feedback. Review of massed concepts occurred in one learning session back-to-back, whereas review of spaced concepts occurred in two separate learning sessions that were two days apart. In both experiments, two days after the last round of review, students took a criterial test that assessed both memory for concept definitions and novel application of these concepts.

Across both experiments, the findings showed that (1) quizzes improved memory for concept definitions relative to restudy, (2) whether quizzes improved novel application of those concepts depended on criterial test's question order and quiz format, and (3) spaced review did not enhance retention or application of the concepts. Below, each finding is discussed in detail; then, the theoretical and educational implications of these findings are outlined, and possible future directions are described.

4.1 Quizzes Improved Memory for Concept Definitions

A large body of research demonstrates that the act of retrieval after learning a set of material improves long-term retention relative to re-exposure to the learning material—a finding referred to as a testing or retrieval practice effect (Roediger & Karpicke, 2006; Rowland, 2014, for recent reviews, see Karpicke, 2017; McDermott, 2021). Retention benefits of quizzing are robust—testing effects are observed across different materials, age groups, initial and criterial test formats, and delays, among other contexts (Dunlosky et al., 2013; Rowland, 2014).

In the current study, participants studied a 38-page packet from a research methods textbook. In both experiments, half of the participants took two definition quizzes on key concepts and received correct-answer feedback after each quiz question (the Test group). Quizzes contained short-answer questions in Experiment 1 and multiple-choice questions in Experiment 2. The remaining half of the participants in both experiments simply read concept definitions (the Restudy group) that served as feedback to the quizzed group. Two days after the second quiz, all participants took a criterial test that asked both definition and application questions.

In both experiments, the Test group outperformed the Restudy group on the criterial test's definition questions, adding to the testing effect literature, specifically with lengthy and complex texts (McDaniel et al., 2015; Uner & Roediger, 2018; Wooldridge et al., 2014). Importantly, both short-answer and multiple-choice quizzes improved retention of concept definitions on a short-answer test. In other words, the occurrence of testing effects did not depend on the match between initial quiz and criterial test formats, in line with a recent meta-analysis that demonstrated significant testing effects despite the mismatch of test formats (Pan & Rickard, 2018). Even after taking a quiz that only required selecting a response (i.e., a multiple-choice

quiz), participants performed better on a later test that required generating a response (i.e., a short-answer criterial test), relative to restudying. Furthermore, the magnitude of testing effects on definition questions was comparable across two experiments, as the Test group performed 7% better than the Restudy group in Experiment 1 (short-answer quizzes) and 9% better in Experiment 2 (multiple-choice quizzes).

One exception to this pattern occurred in Experiment 1, where taking short-answer definition quizzes did not improve memory for concept definitions relative to restudying, but only when definition questions were preceded by application questions on the criterial test. Although the Test group performed 4% better than the Restudy group, this difference was not reliable. A possible explanation could be that the Restudy group had the opportunity to retrieve concept knowledge while answering application questions before the definition questions, thereby increasing the accessibility of those definitions at a later test. This retrieval opportunity, in turn, might have closed the gap between the Test and Restudy groups. In support of this possibility, Restudy participants who first saw application questions did 8% better on definition questions compared to how Restudy participants who saw definition questions first (see Figure 2). This benefit should have been present for the Test group as well, but perhaps the gain from a third retrieval opportunity was less than the gain from a first. However, this was the only one instance in which testing effects were not obtained on definition questions, and in the true test of this effect (i.e., when definition questions came first), testing effects were obtained in both experiments.

One question, of course, is why this pattern did not emerge in Experiment 2. Specifically, in Experiment 2, testing effects were observed on definition questions even when they appeared second on the criterial test. Why would a third retrieval opportunity (i.e., answering application

before definition questions) not help the Test group in Experiment 1, but help them in Experiment 2? In Experiment 1, both the quiz and the criterial test consisted of short-answer questions, whereas in Experiment 2, quizzes contained multiple-choice questions. It is plausible that the change in the question format, from a recognition-based multiple-choice on the quizzes to a generation-based short-answer format on the criterial test, helped the Test group due to variability in re-encoding. As a result, both the Test and Restudy groups benefitted from answering short-answer application questions first in Experiment 2 (8% for the Restudy and 3% for the Test group), and overall testing effects were preserved on the definition questions when they came second on the criterial test.

Finally, considering that the bulk of data collection for Experiment 1 and all data collection for Experiment 2 occurred during a global pandemic and outside of the tight controls of a laboratory setting, this dissertation provides further support for the robustness of testing effects. Participants in both experiments who completed the study online reported various distractions during the different sessions of the experiment, such as being in the same room with other people, loud noises in the background, listening to music, among others. Some participants also reported that they did not pay as much attention or expend as much effort as they would had they been in a laboratory setting. Finally, even though participants were instructed to complete all sessions at the same time when they were sent experiment links, they could complete the sessions at any time they wanted afterward. In fact, although most participants completed sessions 48 hours apart, there was some variability and some participants even completed their sessions very early or very late in the day (e.g., 4 am, 11 pm). Yet, despite these limitations, testing effects were still observed in the current study—reinforcing the robustness of this effect.

4.2 Quiz Benefits on Novel Application Depended on Question Order and Quiz Format

Past research has demonstrated that testing one's memory not only improves retention of what was tested, but can also improve novel application and inference of what was initially learned (Pan & Rickard, 2018). Based on this test-enhanced transfer of learning reported in prior studies, in the current study, participants who took definition quizzes (the Test group) were expected to perform better on novel application questions relative to participants who restudied concept definitions (the Restudy group).

Yet, contrary to this prediction, Experiment 1 showed that short-answer quizzes did not improve novel application relative to restudy when applications questions were answered first on the criterial test. Given that this is the uncontaminated measure of test-enhanced transfer with short-answer quizzes, the lack of a testing effect is surprising, particularly when considering benefits of short-answer quizzing on transfer in prior studies (e.g., Butler, 2010; Hinze & Rapp, 2014; van Eersel et al., 2016, cf. Johnson & Mayer, 2009).

However, short-answer quizzes improved performance on application questions when participants answered definition questions before application questions on the criterial test. In other words, the quizzing benefit on definition questions carried over to application questions that followed. Because remembering concept definitions is presumably helpful in successfully answering novel application questions, and because the Test group correctly recalled more definitions than the Restudy group on the first half of the criterial test, this could explain why testing effects were observed on the application questions. Interestingly, the Test group seems to have benefitted from having answered definition questions prior to application questions, as Test participants who first saw definition questions performed about 7% better on application questions compared to Test participants who first saw application questions (see Figure 2).

Restudy participants who first saw definition questions, on the other hand, performed about 5% worse on application questions relative to Restudy participants who were first given the application questions.

Although Experiment 1 demonstrated benefits of short-answer quizzing on novel application questions when definition questions came before application questions, the same was not true in Experiment 2—no multiple-choice quizzing benefits were observed on application questions. According to Pan and Rickard's meta-analysis (2018), elaborative retrieval practice—defined by broad re-encoding conditions (e.g., explanatory recall, higher- and lower-order quizzes combined) and elaborative feedback (e.g., restudy of all learning material, explanatory feedback)—is associated with greater test-enhanced transfer. It is possible that multiple-choice definition quizzes, which only required participants to select a concept's definition among lures rather than generate information about that concept, and only receiving correct-answer feedback with the definition did not yield elaborative retrieval conditions. In fact, much of the previous research showing testing effects on application- and inference-based criterial tests employed generative tests during retrieval practice (i.e., cued recall or free recall tests) (Pan & Rickard, 2018). Further, the studies that did use multiple-choice initial tests only found test-enhanced transfer when quizzes assessed novel application or inference (e.g., Agarwal, 2019; McDaniel et al., 2015), but not when quizzes assessed factual information (e.g., Agarwal, 2019; McDaniel et al., 2013).

Although short-answer definition quizzes might afford more of the elaborative processes required to improve subsequent transfer, multiple-choice quizzes may need to be scaffolded further to boost transfer. It is possible that with more detailed feedback or a chance to restudy the textbook (or parts of the textbook), participants could have benefitted from definition quizzes on

later novel application questions. Furthermore, including application-based questions on the quiz could have facilitated later novel application performance, because answering application-based questions could have promoted more elaborative processing of concepts than recalling definitions. Further, based on a transfer-appropriate processing framework (Morris et al., 1977), matching the requirements between quiz and criterial test might have resulted in test-enhanced transfer.

4.3 Spaced Review Did Not Enhance Verbatim Retention or Novel Application

Past research has shown that distributing study of to-be-learned material over time rather than placing it in close succession improves long-term retention (Cepeda et al., 2006). Distributed practice effects are robust; they are observed across a variety of learning contexts (e.g., age groups, learning material, see Carpenter, 2017; Cepeda et al., 2006; Dunlosky et al., 2013; Maddox, 2016 for reviews).

In the current study, after reading from a textbook, participants reviewed key concepts from the textbook twice, where one set of concepts was massed and the other set was spaced. When review was massed, participants took a concept definition quiz (the Test group) or restudied concepts (the Restudy group) twice, back-to-back within the first session. When review was spaced, by contrast, participants completed the first quiz or the first restudy in the first session, and the second quiz or the second restudy in the second session, two days later. Regardless of the review timing, the retention interval was two days; participants took a criterial test on massed concepts in the second session and a criterial test on spaced concepts in the third session, with all sessions two days apart.

Based on the robustness of spacing effects, participants should have recalled definitions of concepts that they reviewed spaced better compared to the ones they reviewed massed. However, this was not the case in either experiment. In both Experiments 1 and 2, Restudy and Test groups performed similarly on massed and spaced concepts when examining definition questions on the criterial test. These results are surprising considering the extant literature demonstrating mnemonic benefits of distributed practice. In both Experiments 1 and 2, the Restudy group performed 3% better on definition questions for concepts they reviewed spaced over massed. Though in the expected direction, this difference was not reliable. Performance of the Test group, on the other hand, did not reveal any differences in the expected direction.

Spacing effects were observed in one particular condition in Experiment 2, however; the Restudy group recalled more concept definitions under spaced than massed review only for one set of spaced concepts. Specifically, when concepts in Set 1 were spaced, Restudy participants remembered the definitions of spaced concepts better than definitions of massed concepts. This pattern was also observed in Experiment 1 using logistic regression. As mentioned earlier, definition questions for Set 1 concepts primarily asked for a definition given the concept name, whereas half the definition questions for Set 2 concepts asked for the name of the concept given its definition. Generating definitions may have been more difficult than naming the concept, and Restudy participants may have therefore benefitted more from spaced review on those questions—in line with study-phase retrieval accounts of spacing effects. However, the finding that the presence of spacing effects depended on which set of concepts was spaced was not predicted (which set of concepts was spaced was simply a counterbalancing variable); thus, the above explanation is post-hoc. Furthermore, the fact that the same pattern did not generalize to

the remaining half of the learning material, or to the Test group suggests that this finding might not be reliable.

In addition to the above results, no spacing effects were observed for application questions either, across both Test and Restudy groups. Whether spaced review would improve answers to novel application questions was less clear at the outset based on prior research. However, given the robustness of spacing effects and spacing benefits on related forms of transfer, it was predicted that participants would perform better on application questions for concepts that were reviewed in a spaced fashion. Neither experiment confirmed this prediction; novel application performance for massed and spaced concepts were almost identical across both experiments.

Why were spacing effects not obtained in these experiments? First, it is possible that the way in which massing was implemented in both experiments affected the results. As a reminder, instead of seeing an item back-to-back, participants saw two presentations of the same question or definition nine items apart in the massed condition. This was done to better represent how students review learning material outside of the laboratory. For instance, instructors might allow students to retake practice quizzes, or students might restudy for an exam using a deck of flashcards in the same or similar order. Although the way massing was implemented in the current study aligns more closely with student learning, the small spacing it introduced to the massed review conditions could explain differences between current findings and previous research, given that many of the prior studies contained truly massed conditions. A follow-up experiment currently underway will address this question by using a fully massed condition, where items are presented immediately after each other.

Nonetheless, the difference between reviewing a concept nine items apart (i.e., the massed condition) and two days apart (i.e., the spaced condition) is large, and past research has demonstrated significant lag effects (i.e., retention benefits after a longer than a shorter lag between repetitions). However, although the Restudy group performed 3% better on definition questions after spaced than massed review, the Test group did not even show a trend in the right direction (-3%).

As noted previously, a second possibility for the absence of spacing effects is the relation between the spacing gap of two days and the retention interval of two days. Several spacing effect studies have shown that the magnitude of spacing effects varies when the relation between spacing gap and retention interval varies, and the optimal spacing gap for a given retention interval follows an inverse U-shaped function. That is, for any given retention interval, introducing some spacing and increasing it enhances later retention up until a certain spacing gap, and retention starts decreasing as the spacing gap gets larger beyond the optimal gap (Cepeda et al., 2008; Maddox, 2016). Therefore, although spacing effects were not observed with a gap of two days and a retention interval of two days in this design, changing the retention interval or the spacing gap could yield spacing effects. As one example, if the spacing gap in both experiments was in the order of several hours (i.e., a spacing within the same day) rather than two days, spacing effects could have been observed with a two-day retention interval. Alternatively, if the retention interval was one week rather than two days, spacing effects might have emerged (Cepeda et al., 2008). Indeed, one prior study using simpler materials and a slightly different paradigm has demonstrated mnemonic benefits of longer spacing (two sessions spaced out one day) over shorter spacing (within one experimental session) only when the retention interval was longer (one week), but not when the retention interval was shorter (one

day) (Pyc et al., 2014; for similar findings, see Balota et al., 1989). Because review timing was manipulated within-subjects, a retention interval of one week for a spacing gap of two days would have increased the number of sessions from three to four in both experiments. Given the difficulty of conducting a multi-session online experiment, exploring this possibility awaits future research.

A third possibility regarding why spacing effects were not obtained is the timing of sessions within online data collection. Although each of the three sessions was designed to be 48 hours apart, with the restraints of online data collection, it was not always possible to control the timing of sessions. As a result, some participants completed the sessions less or more than 48 hours apart, though the majority did complete all sessions on the day they were supposed to (collapsed across experiments, the average time between Sessions 1 and 2 was 2.04 days, and the average time between Sessions 2 and 3 was 2.01 days). The variability introduced in how much spacing occurred in the spaced review condition or how long the retention interval was for both massed and spaced review conditions could have masked spacing effects. However, some of Experiment 1's participants completed all three sessions in the laboratory prior to the switch to remote instruction in March 2020. When data from only those participants were examined, the conclusions regarding spacing effects did not change. Furthermore, the time between Sessions 1 and 2 (i.e., the spacing gap for the spaced condition) and the time between Sessions 2 and 3 (i.e., the retention interval for the spaced condition) did not significantly predict criterial test accuracy on spaced items in a simple regression model. This was the case for both experiments and with both in-person and online participants. In other words, the variability in when exactly participants completed the sessions did not matter, at least in the current study, ruling out the possibility that lack of control over session timing minimized spacing effects.

It is also important to consider methodological differences between a typical spacing effect study with simple verbal materials and the current study. In a standard spacing effect experiment, participants might study word pairs three times. These three presentations would either be back-to-back or distributed over time (e.g., by placing other word pairs in between, or spacing out learning sessions over multiple days). That is, the massing or spacing would occur in the initial learning phase. In the current study, however, participants first read a 38-page packet from a textbook, where they were first introduced to the key concepts. After this initial learning session, the spacing manipulation was introduced. In other words, rather than the initial learning being massed or spaced like in most prior spacing research with simple materials, the review of the learning material was massed or spaced in the current study. Thus, participants had some time in between the first time they learned about a concept from the textbook and the first time they reviewed that concept via either testing or restudy.

With lengthy texts, it is difficult to mass or space initial learning; for instance, it may not be realistic for students to reread entire textbook chapters back-to-back. Even if students did read chapters back-to-back, because of the length of the material, there would be some spacing introduced between the first and second exposure to the concepts from chapters. Repeating pages or paragraphs could solve this issue, but this approach would certainly sacrifice the textbook's flow. To make matters more complex, concepts are often presented in a coherent narrative in textbooks rather than in isolated, distinct sections. Specifically for the current study, many of the key concepts were mentioned more than once throughout the packet participants read. As such, participants were exposed to the concepts at least three times (reading the textbook and two concept reviews via testing or restudy), and the timing of only the last two presentations was varied (massed or spaced).

Thus, in some ways, the current study examined how expanding and contracting spacing schedules compare to each other at a delayed test, when the initial spacing is held constant. For concepts reviewed massed, initial spacing between reading a concept in the textbook and reviewing it for the first time was greater than the spacing between the first and second reviews (similar to a contracting schedule). For concepts reviewed spaced, on the other hand, the initial spacing between reading the concept in the textbook and reviewing it the first time was less than the spacing between the first and second reviews (similar to an expanding schedule). In other words, the current study demonstrated that, when the first spacing (between the first and second presentation of an item) was equivalent across conditions, expanding or contracting schedules led to similar subsequent recall—a finding reported in prior literature (e.g., Karpicke & Bauernschmidt, 2011; Toppino et al., 2018, cf. Küpper-Tetzel et al., 2014).

Lastly, consideration of the lack of spacing effects on novel application questions is needed. Despite numerous studies examining spacing effects on verbatim long-term retention, there are only a few studies that investigate whether spaced review (restudy or test) improves novel application or inference (e.g., Rawson & Kintsch, 2005). However, because spacing effects on verbatim retention were not obtained in the current study, the current findings cannot provide conclusive evidence regarding spacing effects on novel application questions. That is, spaced review may improve novel applications in a scenario in which spaced review improves verbatim retention.

4.4 Theoretical Implications

Across both experiments, the absence of spacing effects on both kinds of questions and the absence of stable testing effects on application questions could well be due to the particular methodologies used; however, given the authentic learning contexts simulated in the current

study, these findings provide limitations on the applicability of current theories of spacing and testing.

In particular, it is surprising that short-answer quizzes did not improve novel application performance on the criterial test when application questions were presented first. This finding fails to replicate past research demonstrating transfer to application or inference questions after initial short-answer tests (e.g., Butler, 2010). One interpretation of this discrepancy relates to the complexity in material and the variability in the kinds of application questions used. It is possible that the current study identifies a boundary condition of retrieval practice as a tool to improve novel applications. That is, when students learn from a lengthy and complex chapter, when they are quizzed on definitions of key concepts with corrective definition feedback, and when they receive questions that assess a variety of ways to apply different concepts, testing may not help with novel applications.

Furthermore, the lack of benefits of spaced over massed review in the current study is difficult to reconcile with the existing literature on spacing effects and lag effects (Carpenter, 2017; Cepeda et al., 2006). Given that the current study employed much longer and more complex materials than those used in prior studies, and also given that the review of the learning material rather than initial learning was spaced or massed, findings discussed above provide possible limitations to the benefits of distributed practice.

The current findings have some implications for the encoding or context variability accounts of distributed practice effects. According to these accounts, as the time between the repetitions of the learning material increase, the more the context in which the material is studied changes. These different contexts, in turn, provide cues to retrieve the material at a later point, relative to remembering the learning material studied close in time with more similar contexts.

Although the current study did not demonstrate any benefits of spaced review over massed review, these findings could be due to the complexity of the material. With lengthy and complex materials such as a textbook chapter, the material itself may have some context due to the narrative within the chapter and the relatedness of the concepts explained in it. As such, varying the internal or external contexts (e.g., mood physical, environment) might not confer additional mnemonic benefits for such materials. Thus, even though the variability in context might aid simpler verbal materials (e.g., word lists) and thus could be the mechanism with which spacing effects occur, complex materials such as textbooks or other educationally-relevant materials might not benefit from variability in encoding or context—providing nuance to the encoding variability account.

Because spacing effects were not obtained on definition questions in either experiment, the current study could not examine the possible combined benefits of spaced testing on application questions. As mentioned in the Introduction, current theories of spacing and testing have overlap in the mechanisms proposed to underlie these effects. Given that both study strategies are considered to increase the accessibility of information at a later point, which might be necessary but not sufficient to improve transfer of learning, combining spacing and testing might not improve novel applications above and beyond one strategy alone. Thus, the question of whether spacing and testing benefits on transfer are interactive is still theoretically relevant, and this issue should be addressed within studies that demonstrate separate benefits of each study strategy on transfer.

Finally, because of the methodological differences between the current study and prior research (particularly on spacing effects) with which theories of distributed practice and retrieval practice were proposed, interpreting the current findings with extant theoretical perspectives is

challenging. However, these methodological differences also form aspects of the current study that allow findings to be more readily applicable to classroom learning. Thus, although current theories cannot easily accommodate findings reported above, the current study has significant educational implications.

4.5 Educational Implications and Future Studies

Students are expected to not only remember what they learned in textbooks or lectures, but they are also expected to use this knowledge in novel ways on exams and outside the classroom—a critical goal of education. This dissertation provides a paradigm in which the effectiveness of retrieval practice and distributed practice can be examined on both verbatim retention and novel application. The question of interest was not how students should first be introduced to a set of material; rather, the question was how students should review the material they had previously encountered. Thus, the paradigm simulates a situation in which students first learn about a concept in class or within a textbook that they then review prior to an exam. The question examined in the current study was whether combining spacing and testing, when students reviewed key concepts from the initial learning material, improved their ability to not only remember concept definitions but also to apply concept knowledge to novel questions.

Taking definition quizzes improved retention of those definitions across both experiments, relative to rereading the definitions. Based on this finding, students should test themselves on material they previously learned and seek corrective feedback. Fortunately, the way in which students test themselves may not be as important, as the benefit of taking practice quizzes on later criterial tests was observed regardless of the match between question formats.

In this paradigm, the timing of these quizzes or of restudy did not matter either; students performed similarly on definition questions after reviewing concepts massed or spaced. If

students have a test in two days, spacing out review across two days or doing it all in one sitting may not matter. Nonetheless, this recommendation is not set in stone; changing the retention interval (i.e., how soon after the test is from the last time material was reviewed) or the spacing gap (i.e., how much time there is until the same material is reviewed again) would likely paint a different picture.

The current study presents challenges to the retrieval practice and distributed practice literatures, as no robust benefits of either strategy were observed on novel application questions. Though these findings could be taken as evidence that strategies that reliably improve students' verbatim retention are not effective when students' goal is to improve their mastery of concept knowledge, further studies are needed before firm conclusions can be made. Future studies should first examine whether spacing effects extend to comprehension from lengthy and complex texts when a particular spacing gap and retention interval lead to verbatim retention benefits. Without obtaining any spacing effects on the definition questions in this paradigm, it is not fair to discard spacing as a strategy that could be used to improve novel application and inference. In addition, in the current study, the timing of review was manipulated within participants, where half of the material was reviewed massed and the other half was reviewed spaced. However, it is likely the case that students study for their exams either in massed or spaced fashion, rather than split the material to review only some of it massed or only some of it spaced. Although review timing was manipulated within participants to increase the power to detect a spacing effect in this paradigm (if it exists), a more educationally relevant extension of the findings would demonstrate spacing benefits when review timing is varied between participants.

Furthermore, regarding testing effects on application questions, future research should examine the role of application- and inference-based questions when given on practice quizzes.

When students are quizzed with such higher-order questions, they may better anticipate the kinds of ways they need to use concept knowledge, and they might be better equipped to tackle novel higher-order questions on an upcoming exam.

The primary goal of the current study was to examine whether the combined benefits of testing and spacing on verbatim retention would extend to novel application. Because spacing effects were not observed on the definition questions, it was not possible to assess the generalizability of spaced testing to transfer. Thus, this question remains to be answered in future research.

Finally, it is worth noting that student learning was examined under relatively authentic conditions in this dissertation. Students learned from the kinds of text material they are typically presented within college courses, they learned and studied information across multiple learning sessions distributed over the course of a week, and they were assessed on both verbatim memory and transfer. Furthermore, there also was variability in item difficulty; some concepts (e.g., *independent variable*) were fairly easy for most participants, some (e.g., *error variance*) were much more difficult to most, and some concepts were somewhere in between. Similarly, participants were given a variety of question stems for both definition and application questions; some questions asked for a concept definition given the concept name, some questions asked participants to identify the concept given a brief scenario, some questions asked participants to generate an example demonstrating a particular concept, to describe a few. Although these aspects of the design might have minimized the magnitude of observed effects, they represent a much more authentic learning situation. Relatedly, due to online data collection resulting from the pandemic, participants completed the experiment in their own environment, at many different times of the day, with many more distractions than a laboratory setting would present, among

other considerations. These issues undoubtedly introduced more error variance, thus minimizing the ability to detect effects. Nonetheless, these experiments represent a much more realistic portrayal of student learning.

Altogether, the current study presents a paradigm through which the effectiveness of study strategies can be examined on different learning outcomes. By incorporating the considerations mentioned above, it will become clearer under what conditions spacing and testing improve students' memory and comprehension of lengthy and complex texts.

References

1. Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.
2. Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning?. *Journal of Educational Psychology*, *111*(2), 189.
3. Anderson, F. T., & McDaniel, M. A. (2021). Restudying with the quiz in hand: When correct-answer feedback is no better than minimal feedback. *Journal of Applied Research in Memory and Cognition*.
4. Appleton-Knapp, S. L., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research*, *32*(2), 266-276.
5. Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*(5), 316-321.
6. Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566-577.
7. Balota, D.A., Duchek, J.M., & Logan, J.M. (in 2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J.S. Nairne (Ed.), The foundations of remembering: Essays in honor of Henry L. Roediger, III. New York: Psychology Press.
8. Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3-9.
9. Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger III, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*, *21*(1), 19.
10. Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612-637.
11. Barnett, J. E., & Seefeldt, R. W. (1989). Read something once, why read it again?: Repetitive reading and recall. *Journal of Reading Behavior*, *21*(4), 351-360.
12. Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen, R. H., Singman, H., . . . Fox, J. (2019). Linear mixed-effects models using 'Eigen' and S4. (lme4 version 1.1-21) [Computer software]. Retrieved from https://github.com/lme4/lme4/
13. Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective?. *Cognitive Psychology*, *61*(3), 228-247.

14. Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392-402.
15. Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849.
16. Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470–497). New York, NY: Cambridge University Press.
17. Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts, and categories: Studies in cognition* (pp. 408-437). Cambridge, MA: MIT Press.
18. Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118.
19. Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied*, *23*(4), 433.
20. Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290-298.
21. Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review*, *26*(2), 331-340.
22. Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30-41.
23. Carpenter, S. K. (2017). Spacing effects in learning and memory. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (J. H. Byrne, Series Ed.), pp. 465-485. Oxford: Academic Press.
24. Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619-636.
25. Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268-276.
26. Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671-682.
27. Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474-478.

28. Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*(6), 760-771.
29. Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633-642.
30. Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481-495.
31. Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: what can sequencing effects tell us about inductive category learning?. *Frontiers in Psychology*, *6*, 505.
32. Carvalho, P. F., & Goldstone, R. L. (2019). When does interleaving practice improve learning? In J. Dunlosky & K. A. Rawson (Eds.), *Cambridge Handbook of Cognition and Education* (pp. 411–436). New York: Cambridge University Press.
33. Cassady, J. C. (2000). Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research, and Evaluation*, *7*(1), 12.
34. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354.
35. Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095-1102.
36. Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*(3), 215-235.
37. Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153-176.
38. Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In *Psychology of learning and motivation* (Vol. 53, pp. 63-147). Academic Press.
39. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.
40. Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications. (Original work published 1885).
41. Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215-228.
42. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

43. Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717-741.
44. Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition*, 1-14.
45. Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology*, *60*(7), 991-1004.
46. Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning* (pp. 9-46). New York: Academic.
47. Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729-757.
48. Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93-103.
49. Heiman, G. W. (2002). *Research methods in psychology* (3rd ed.). Boston, MA: Houghton Mifflin.
50. Hintzman, D. L., & Rogers, M. K. (1973). Spacing effects in picture memory. *Memory & Cognition*, *1*(4), 430-434.
51. Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*(4), 597-606.
52. Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. (2016). Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge. *Educational Psychology Review*, *28*(4), 853-873.
53. Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441.
54. Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621.
55. Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 12-19.
56. Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998-1005.
57. Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

58. Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97-103.
59. Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction*, *36*, 38-45.
60. Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469.
61. Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (J. H. Byrne, Series Ed.).
62. Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250.
63. Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775.
64. Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237-284). Academic Press.
65. Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704.
66. Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966-968.
67. Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science*, *19*(6), 585-592.
68. Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337-1344.
69. Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. (2019). How the amount and spacing of retrieval practice affect the short-and long-rerm retention of mathematics knowledge. *Educational Psychology Review*, 1-19.
70. Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, *28*(6), 684-706.
71. McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention?. *Journal of Experimental Psychology: Applied*, *21*(4), 370-382.
72. McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1417.

73. McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516-522.
74. McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360-372.
75. McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18-26.
76. McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, *72*, 609-633.
77. McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3.
78. Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*(3), 390-407.
79. Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.
80. Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710.
81. Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
82. Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwait (Eds.), *International encyclopedia of education, 2nd ed.* (Vol. 11, pp. 6452–6457). Oxford, England: Pergamon Press.
83. R Development Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org
84. Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, *27*(3), 431-452.
85. Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?. *Journal of Experimental Psychology: General*, *140*(3), 283.
86. Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*(6), 1004-1010.

87. Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70.
88. Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27.
89. Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.
90. Roediger III, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248.
91. Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*(9), 1209-1224.
92. Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*(6), 481-498.
93. Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.
94. Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 755.
95. Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology, 51*(1), 123-138.
96. Tse, C. S., Balota, D. A., & Roediger III, H. L. (2010). The benefits and costs of repeated testing on the learning of face–name pairs in healthy older adults. *Psychology and Aging, 25*(4), 833.
97. Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 381-391.
98. Uner, O., & Roediger III, H. L. (2018). The effect of question placement on learning from textbook chapters. *Journal of Applied Research in Memory and Cognition, 7*(1), 116-122.
99. Wheeler, M. A., & Roediger III, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*(4), 240-246.
100. Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*(3), 214-221.
101. Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*(1), 16-27.

Appendix A

Concept Sets

I created two concept sets such that concepts within a set were more similar to each other and less similar to those in the other set. Specifically, I used data from a prior study using the same learning materials (Anderson & McDaniel, 2021) and conducted a principal component analysis to examine which items grouped together. Using that as a starting point, I made qualitative judgements based on conceptual similarity or dissimilarity. For instance, I made sure to place concepts presented in the same section of the textbook, such as content and construct validity, in the same concept set.

Set 1: conceptual replication, confounding variables, construct validity, content validity, external validity, internal validity, random sampling, reliability, subject mortality, volunteer bias

Set 2: balancing, conditions, dependent variable, error variance, independent variable, operational definition, order effects, random assignment, repeated-measures design, state and trait characteristics

Appendix B

Definition and Application Questions for Concepts

1-Balancing
Definition: In the context of an experiment, what is balancing participants?
Application: A researcher is studying the effects of music on reading comprehension. Participants will either listen to classical music, jazz music, rock music, or no music while reading a passage and then they will answer a series of questions. He wants to control for reading level, so he pretests reading level. He then puts one out of the top four readers into each condition and so on. What type of design is he using to control for reading ability?

2-Conceptual Replication
Definition: What is a conceptual replication?
Application: One study showed that male participants rated a female research assistant as more attractive after they crossed a very high suspension bridge than before they crossed the bridge. The researchers argued that the male participants misattributed their increased heart rate to their attraction towards the research assistant. What is one way you can conceptually replicate this finding?

3-Conditions
Definition: What is the difference between a control condition and the other conditions used in an experiment?
Application: In your study, you create different conditions by playing different types of music to participants. After a while, you suddenly pull out and shoot a (blank) pistol. You then measure participants' anxiety level to determine whether different types of music cause people to remain more or less calm in the face of startling stimuli. Name 3 possible conditions and identify one as a control condition.

4-Confounding variables
Definition: What are confounding variables?
Application: A researcher conducts an experiment on participants' memory of a story, comparing the recall of those who read it silently to that of others who read it out loud, hypothesizing that reading out loud will improve recall. Identify a potential confounding variable.

5-Construct validity
Definition: What is construct validity?
Application: A researcher asks people how many nights per week they spend time with friends and uses this to assess an individual's sociability. Another researcher argues that sociability is better determined by the number of friends a person has. The second researcher is criticizing the _____ of the first researcher's measurement.

6-Content validity
Definition: What is content validity?

Application: A therapist is studying a group of children with ADHD. She is classifying the children as having mild, moderate, or severe ADHD by counting the number of disruptions they cause in one hour of class time at school. Girls with ADHD tend to exhibit less disruptive symptoms than boys. For instance, they are more likely to appear spacey or disorganized. Does the therapist's measurement of ADHD have content validity? Why?

7-Dependent variable
Definition: We conduct an experiment to see if our manipulations result in a change to the _____.
Application: Generate a basic research study and identify the dependent variable.

8-Error variance
Definition: What is error variance?
Application: A researcher is interested in whether font type influences memory for words. He assigns participants signed up to the morning sessions to one font type, and assigns those signed up to the afternoon sessions the other font type. His labmate tells the researcher that he should randomly assign participants to conditions rather than make assignments based on when they show up. Would you expect a smaller error variance in the researcher's version of the experiment or his labmate's version? Why?

9-External validity
Definition: What is external validity?
Application: You conduct a study examining the effects of parents' education level on the intelligence of your research methods class. What limitations on external validity might this produce?

10-Independent variable
Definition: The _____ is what causes a systematic change in behavior within an experiment.
Application: A professor is interested in whether students' happiness is related to the reward they get for attending class. In the middle of both sections of her class, she gives half of the students a few pieces of candy, and the other half of the students a couple dollars. She measures students' happiness using a questionnaire at the very beginning and the end of class. What is the independent variable?

11-Internal validity
Definition: What is internal validity?
Application: You conduct a study that measures the differences in optimism between 20-year-old men who are in college and 20-year-old men who are not in college. You conclude that being in college causes higher levels of optimism. Does your study have high or low internal validity? Why?

12-Operational definition
Definition: The specific description of how a variable is measured and/or manipulated is called a(n) _____.

Application: You want to conduct an experiment about how background noise influences learning. How would you operationally define the variables in your experiment?

13-Order effects
Definition: _____ is/are a contaminating effect where people "carry-over" their biases/experiences from one part of the experiment to the next.
Application: An experimenter is attempting to determine what type of coffee people prefer best when it is either hot outside or cold outside: light roast, dark roast, or hazelnut. To do so, she collects a sample of participants and has each person try all of the different types of coffee and pick which one they like the best. Some participants do this on a cold day, and those in another condition do this on a hot day. What should she do to prevent order effects in her design?

14-Random assignment:
Definition: Researchers use _____ to produce a balanced, representative sample in each condition.
Application: You want to determine whether or not a new drug can boost understanding of prose passages. You randomly assign participants to two conditions; one in which participants take the drug and study a book chapter, and in the other condition participants take a placebo (sugar pill) and study the same book chapter. They are later tested on the reading. As a researcher, you are under time pressure and are only able to get 15 participants in each condition. Are you confident in the effectiveness of your random assignment? Why?

15-Random sampling
Definition: What is random sampling?
Application: How would you appropriately select a random sample of elementary school students in the St. Louis region?

16-Reliability
Definition: What is reliability?
Application: A student complains to her professor that her essay makes the same points as her friend's but she got a lower grade than her friend. She is complaining that the grading lacks _____.

17-Repeated-measures design
Definition: What is a repeated-measures design?
Application: You are studying whether biking or running is more effective at reducing an individual's amount of body fat. Each participant has their body fat measured before the study. They then bike every day for two weeks and have their body fat measured again. Finally, they run every day for two weeks and have their body fat measured at the end. What type of design are you using?

18-State and trait characteristics
Definition: What is the difference between a state characteristic and a trait characteristic?
Application: Give an example of a situation where extraversion is a state characteristic and an example of a situation where extraversion is a trait characteristic.

19-Subject mortality
Definition: When participants drop out of a study before it is completed it is called _____.
Application: A study measuring whether personality is stable across the lifespan surveys participants once every ten years starting at the age of 10. How might the researchers ethically reduce subject mortality?

20-Volunteer bias
Definition: What is volunteer bias?
Application: An experimenter recruited participants through online advertisements and had participants register online. Why might a problem with volunteer bias arise? How could she have lowered volunteer bias?

Appendix C

Marginally Significant Three-Way and Four-Way Interactions in Experiment 2

The three-way interaction occurred between review type, question type, and spaced concept set, $F(1, 104) = 2.86$, $p = 0.09$, $\eta_p^2 = 0.03$. However, because which concept set was spaced or massed should have no bearing on the magnitude of testing effects for definition and application questions, it is difficult to interpret this interaction. The pattern follows the interaction between question type and review type explained earlier, where testing effects are observed on definition, but not on application questions.

A marginally significant four-way interaction occurred between review type, question type, review timing, and question order, $F(1, 104) = 2.86$, $p = 0.09$, $\eta_p^2 = 0.03$. Pairwise comparisons examining testing effects—separately for massed and spaced concepts, for definition and application questions, and for the two question orders—revealed that although testing effects were not observed on application questions under any condition, testing effects were observed on definition questions under all but one condition. Specifically, the Test and Restudy groups performed similarly on the definition questions that were reviewed massed, when application questions preceded the definition questions ($p = 0.44$). It is possible that taking a previous test with application questions before the definition questions eliminated the benefit of quizzing that occurred during prior sessions. However, the interaction was only marginally significant, and thus these findings may not be stable.

Two of the remaining marginally significant four-way interactions include the two counterbalancing variables (first interaction between review type, question type, spaced concept set, and question order; second interaction between review timing, question type, spaced concept set, and question order), and therefore are difficult to interpret.